Reciprocity at the Workplace

Evidence from Experimental, Personnel and Operational Matched Data

Dissertation

for the Faculty of Economics, Business Administration and Information Technology of the University of Zurich

to achieve the title of
Doctor of Economics

presented by

Beat Hedinger
from Zurich

approved at the request of

Prof. Dr. Uschi Backes-Gellner
PD Dr. Thomas Zwick

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich herewith permits the publication of the aforementioned dissertation without expressing any opinion on its views.

Zurich, December 5, 2007

The Dean: Prof. Dr. H. P. Wehrli

Acknowledgments

First I would like to thank my parents who supported me for more than twenty years to make it possible that I was in the position at all to start this project. Then I would like to gratefully acknowledge the support of my former direct superior, Commander of all Swiss Air Force professional pilots (Ueberwachungsgeschwader UeG), Colonel Bruno Morgenthaler. If it would not have been for his patronage I would never have been able to retrieve the data for this work. Moreover he permitted to reduce my workload at the Air Force to 80% during a shortage of qualified pilots, so that I may dedicate the other 20% writing these pages. No less would I like to thank my Professor, Mrs. Uschi Backes-Gellner, who allowed me to be an external researcher and encouraged me with many ideas on how to proceed with this project. Also I would like to thank Lorenz Goette and his crew for helping to perform the experiments and for further useful advice. Then I am grateful to Thomas Zwick, Stephan Veen, Johannes Mure, Hannah Scott and Jeremy Faux for their helpful comments and interesting opinions. I furthermore much appreciated the computer skills of Jan Stauffacher, best friend since my schooldays. I am also much obliged to my most special friend, Paul Mauguin, who even traveled halfway around the world trying to make sense of these pages! Also important was the operational support of Markus Burkhard during one of my tours in Bosnia that allowed me to spend quite some time on this work. Special thanks to Claudio Noto, who not only encouraged me to start, proceed and finish this work, but was always there to listen to and comment on my peculiar ideas even when flying to holiday-destinations. But most especially, I wish to thank Martina, my wife, who was, despite the huge time-consuming, mind- and conversation-capturing properties of these pages, ready and (still) willing to marry me on the 14th of July this year. I will always be grateful for her love, trust and belief in me, which I will try to reciprocate with all my heart for the rest of my life.

Table of Contents

Figures and Tables

CHAPTER 1:

Introduction

Contrary to the assumptions of neoclassical game theory, recent research found reciprocal preferences to matter in laboratory experiments. With this work I want to contribute evidence that such experimentally found preferences are in fact applicable to the real world of *work*. To do this, I merge experimental data with personnel and operational data of the same employees.

A basic and fundamental assumption of neoclassical economic game theory is that every player is acting according to strictly selfish preferences.[1] This simple assumption helps us to understand economic interrelations, like the competition between firms, the price mechanism of goods and labour markets. However, for many years, researchers voiced doubts about its general applicability to all aspects of the economy.[2] These doubts where not just based on ethical considerations but also on experimentally observed behaviour: People tend to reward the kind and punish the unkind actions of other people even if no future economic benefit can be achieved. Famous experimental examples that detect such behaviour are gift-exchange and ultimatum games:[3] In the gift exchange game, Player One may give money to Player Two and the transfer is doubled by the experimenter. Player Two may then return money to Player One and the return-gift is doubled again. Even though a strictly selfish Player Two would never return a gift, many players do so, which has been interpreted as a clear sign of positive reciprocal preferences. In the ultimatum game Player One has to divide a pot of money. Thereafter Player Two can either accept or reject the division. If he accepts, both players

[1] Already Adam Smith (1776) wrote: "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages."

[2] For an overview see e.g. Fehr and Schmidt (1999).

[3] See e.g. Falk and Fischbacher (2003).

receive their shares, if not both players earn nothing. Even though it would be rational for Player Two to follow strictly selfish preferences and accept any division, many unfair divisions are rejected, which has been interpreted as a clear sign of negative reciprocal preferences.

As a consequence, many researchers adapted economic theories to match these experimental observations.[4] As a next step it is all the more important to find out to what extent these recent theoretical adoptions are applicable to the economic world outside the laboratory. Contributing such evidence is the goal of this work.

As the "economically odd behaviour" mostly appears when personal contacts in prolonged relationships are observed (Gachter and Fehr, 1999; Goette, Huffman and Meier, 2006), I focus on internal labour market relations, in which employers and employees interact over time. The important question is whether positive and negative reciprocal preferences play a role in this relationship and whether experiments can predict such behaviour.

But researching reciprocal behaviour outside the laboratory is difficult, because concerns for reputation and repeated interactions cloud positive and negative reciprocal preferences:[5] When workers reward high wages by high effort levels, this seems like positive reciprocity. But this behaviour may also be motivated by purely selfish preferences: When nobody wants to employ somebody who is known for taking advantage of others, employees tend to mimic positive reciprocal behaviour in order to keep their job.[6]

For the case of negative reciprocity, repeated interactions are just as crucial: Even if one might perceive a strike to be motivated by negative reciprocity, especially if it is costly for employees with piece-rate contracts, these employees may just follow strictly selfish preferences for better wages in the long run.

To solve the problem of indistinguishable selfish and reciprocal preferences I tread a new path of research: I start in chapter two by conducting experiments with employees. In chapter three I then empirically analyse the reciprocal behaviour of the same employees in a real employer-employee situation where reputation may not be at stake. Finally, in chapter four, I combine the experimental data with the empirical data. Doing so I validate if in fact reciprocal

[4] For an overview see e.g. Sobel (2005).

[5] See e.g. Fehr and Schmidt (1999), Fehr and Gachter (2000a), Brown, Falk and Fehr (2002) or Camerer and Fehr (2002).

[6] For an early version of these thoughts see e.g. Shapiro and Stiglitz (1984).

behaviour plays a role in the relationship of employers and employees and whether a laboratory experiment is able to predict real-life behaviour.[7]

By performing experiments with employees in chapter two, I not only gather the crucial data for the final analysis of chapter four, but I also research whether an employer is willing and able to recruit, train or socialize its workforce in regard to reciprocity to match his particular work environment. The more that is at stake and the harder it is to supervise a workforce, the more beneficial positive reciprocity becomes for employees to efficiently work as a team. If this is the case, reciprocity may be relevant for recruiting, training and socialization of employees. For such evidence I focus on a special work environment, which is highly team-sensitive and difficult to supervise in all its aspects. Being a Swiss Air Force pilot myself, I was permitted to conduct a typical laboratory experiment during a compulsory training workshop for all pilots. The analysis of this experimental data shows higher levels of positive reciprocity among pilots compared to the standard student population. According to further empirical analyses, it even seems highly probable that this special workforce is much more positive reciprocal than the average population. This difference prevails even if the pilots play against outsiders. This interesting finding shows that either the recruiting, training or socialization of pilots achieved the goal of finally employing highly positive reciprocal team-players. I derive that reciprocity is important for certain employers influencing their recruitment, training and socialization policies.

To gather more detailed information about the applicability of reciprocity to the labour market, in chapter three I search for a real life employer-employee situation, in which only reciprocity, and not reputation, matters. In order to find such a situation, I use two unique datasets. All pilots of the Swiss Air Force are evaluated annually upon a subjective performance evaluation system called "LOBE". Since this system is based on *subjective* performance evaluations, the concerned pilots may regard any improved or deteriorated evaluation as a friendly or unfriendly action of a superior. As this might induce a respective reciprocal reaction, I look for an activity that would not be relevant for considerations of evaluation or reputation. Fortunately there is additional operational data that is unobserved by the evaluating superiors due to military regulations. And indeed, the pilots seem to choose their unobserved effort levels according to changes in their subjective performance

[7] Even though the findings of chapter two and three are important for chapter four, all three chapters are written as individual complete research articles.

evaluations. Hence I am able to show that reciprocity matters for the performance of unobserved tasks.

As this apparent reciprocal behaviour may be very important to many employers, the next question is, whether it can in fact be associated to reciprocity. To answer that, I generate a unique dataset combining not only the subjective performance evaluations with unobserved operational data, but also with experimental data for the same individuals. Therefore in chapter four I am able to apply a new research method that merges the three datasets to asses whether the experimentally assessed individual reciprocity of chapter two predicts the individual reciprocal behaviour I empirically find in chapter three. And indeed, the experimental findings are able to predict the levels to which the pilots are reciprocating positive and negative changes in their subjective performance evaluations. In my opinion, these findings born out of such a unique combined dataset may not only be essential for experimental researchers, but equally important for employers to improve the efficiency of workplace relations. I thus summarize my findings in chapter five and indicate possible implications.

CHAPTER 2:

Cooperation in the Cockpit

Experimental Evidence of Reciprocity and Trust among Swiss Air Force Pilots

2.1. Introduction

Voluntary cooperation is important in organisations due to the incomplete nature of contracts. Much attention has been devoted to the cooperation between firms and workers (Prendergast, 1999; Kandel and Lazear, 1992; La Porta et al., 1997; Akerlof and Kranton, 2005). In this case the typical problem involves the difficulty to enforce complex contracts to render a firm efficient. This issue can often be resolved through repeated interactions, so that a worker's incentive to comply is to profit from future interactions.[8] If, in addition to that, a minimal fraction of subjects is not purely selfishly motivated, cooperation is enhanced and sometimes even possible in one-shot interactions (Fehr and Schmidt, 1999; Brown, Falk and Fehr, 2002).[9]

In many cases, cooperation between a worker and his fellow worker is just as important for a firm as is its efficiency (La Porta et al. 1997). This interaction differs in two important aspects from the interaction between a worker and a firm:

[8] The effects of reputation can prove efficient (Fehr and Schmidt, 1999; Fehr and Gachter, 2000a; Brown, Falk and Fehr, 2002; Camerer and Fehr, 2002).

[9] Non-purely selfish individuals are commonly modelled by utility functions that include other individuals' actions and payoffs (see e.g. Levine (1998), Fehr and Schmidt (1999), Falk and Fischbacher (2000), Bolton and Ockenfels (2000) or Charness and Rabin (2002)).

First, the interaction is harder to observe for the firm: It is unlikely to be enforced by incentives such as those created through repeated interactions between workers and firms (La Porta et al., 1997). Since employees often refrain from reporting an incident to the firm, where fellow workers do not cooperate, a worker cannot be disciplined by the fear of loosing his job. Hence it is difficult for the employer to establish cooperation among workers.

Second, social preferences may be more powerful in interactions between workers, because workers mostly interact with fellow workers. Therefore they might identify more strongly with one another than for example with their principal (Akerlof and Kranton, 2003).

It follows that firms have to rely on social preferences of their employees to ensure the cooperation among their workers. These social preferences can be expressed through positive and negative reciprocal behaviour. If such reciprocity is anticipated by other workers, it can lead to trust and the cooperation of a whole group (Fehr and Gachter, 2000b; Falk and Fischbacher, 2000; La Porta et al., 1997; Carpenter, Burks and Verhoogen, 2005).

In this chapter, I measure positive and negative reciprocity in a group of subjects for which cooperation at work is literally a question of life or death: I conducted a moonlighting game with pilots of the Swiss Air Force. While previous studies have shown that workers cooperate more than students (Fehr and List, 2004; Carpenter, Burks and Verhoogen, 2005), I examine two novel issues: First I study preferences for positive versus negative reciprocity in workers when negative reciprocity can have very high costs to the principal. Second I examine the robustness of reciprocity to social distance when cooperation is indispensable.

Starting with the first issue, I argue that in some cases negative reciprocity may bear more adverse effects than the efficiency gains achieved by enforced cooperation. In the context of aviation, for instance, there are spectacular examples of how negative reciprocity can be costly to an organisation: On October 28, 2003, an airliner flew from Leipzig to Zurich. After an uneventful smooth flight, suddenly the captain decided to perform the duties of the co-pilot in addition to his own. The co-pilot stopped doing his duties almost altogether, because he felt himself left out. Consequently, he did not use the proper map to monitor the captain's approach nor did he check the visual references the captain claimed to see on final approach. In heavy fog, the aircraft crashed next to the runway due to the captain's bad judgment. A subsequent investigation ruled out any technical failure (AAIB, 2006). Although the aircraft was a total loss, on account of pure luck, no one was hurt.

It can be argued that negative reciprocity played an important role in the occurrence of the accident: The co-pilot was unwilling to perform his duties after having been left out by the

captain, even though it would have been especially important to assist a captain, who is performing the duties of two crewmembers at the same time. Hence it is particularly interesting to study if the vital cooperation among pilots is typically established through higher levels of positive reciprocity rather than higher levels of negative reciprocity.

Focusing on my second issue, it can be argued that a further feature may have contributed to the detriment in my example: that the captain relied on his rank to relieve the co-pilot of his duties. Any identification process between the pilots, which could have lead to stronger positive and weaker negative reciprocity, would have been destroyed. Therefore social distance, in the form of hierarchy, may further hinder cooperation in such an environment.[10] As cooperation is essential among pilots to remain effective, neither negative reciprocity nor social distance should be a critical factor. Hence it is also interesting to study if social distance, particularly in the form of hierarchy, influences levels of reciprocity.

To research the first issue, I compare Air Force pilots and students playing the moonlighting game for real money. I find that positive reciprocity is indeed stronger than negative reciprocity among pilots compared to students. In addition trust is shown to be dependent on the level of expected reciprocity.

To research the second issue, I first compare positive and negative reciprocity of pilots paired with higher-ranking pilots to positive and negative reciprocity of pilots paired with pilots of equal rank. I do not find any effect of perceived social hierarchy on actions and reactions of pilots and no statistically significant difference in the levels of positive and negative reciprocities was detected.

To further stress the second issue, I test amongst the pilots the general assumption that high levels of cooperation among workers are based on a higher degree of group identity similar to the studies of Gachter and Fehr (1999) and Goette, Huffman and Meier (2006). The pilots were paired with students to investigate pilots' behaviour towards total outsiders. Surprisingly positive reciprocity does not vanish.

My results suggest differences in basic social preferences between pilots and students. It seems tempting to conclude that these differences are achieved through screening, training or socialization of the Swiss Air Force.

[10] In addition to that, it has been shown that perceived social hierarchy may hinder the optimal performance of subjects (Hoff and Pandey, 2006).

The remainder of this chapter is organized as follows: In section two I provide some details about Swiss Air Force pilots. Section three discusses the experimental design and provides the predictions. Results are discussed in section four. Finally section five summarizes.

2.2. Institutional Background

In order to better understand the screening, training and socialization of Swiss Air Force pilots, this section provides background details about their careers.

Every year the Swiss Air Force picks 6 to12 potential pilots out of hundreds of twenty-year-old applicants, who, among other criteria, are chosen for their teamwork capabilities (Noser, 2003; Airforcepilot, 2004).

To further train cooperation, the applicants have to attend 56 lessons and 3 years of practical education in "Crew Resource Management" during pilot school.[11] This particular course has been developed on the basis of management training and aims to optimise the use of human resources in the cockpit (Helmreich, Merritt and Wilhelm, 1999).[12] After pilot school, every flying officer is required to attend to refresher courses on a regular basis.

In their work environment, pilots become more socialized because of the heavy dependence on each other: Large helicopters are always crewed by two pilots due to their complexity. Jet fighters never fly alone to increase efficiency and firepower. A single pilot is unlikely to succeed in accomplishing an actual task especially when confronted with an emergency situation. Therefore crewmembers must fully cooperate in order to minimize the probability of loss of human lives and best accomplish their task.

For all these circumstances, I view Swiss Air Force squadrons as being a work environment where cooperation has high efficiency gains and where screening, training and socialization is used to enhance such cooperation.

[11] The Commander of Screening & pilot basic training Swiss Air Force provided this information.

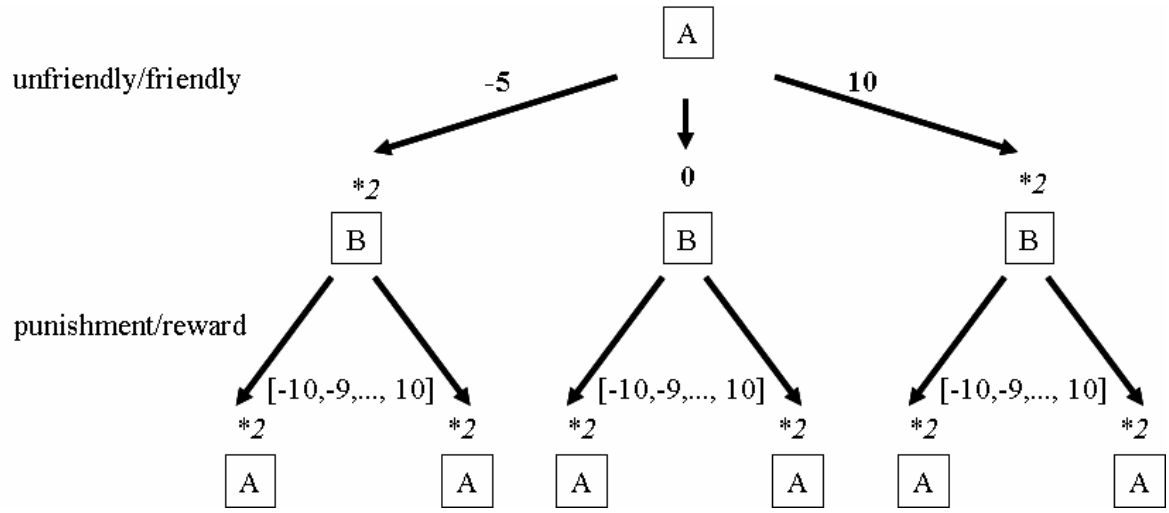
[12] It enhances communication, teamwork, situational awareness, decision-making capabilities, leadership and stress management (Helmreich, Merritt and Wilhelm, 1999).

2.3. Experimental Design

2.3.1 The Game

The subjects participated in several moon lighting games (Abbink, Irlenbusch and Renner, 2000). In this particular game, Player A has the choice of being friendly, neutral or unfriendly to Player B. Player B in turn rewards or punishes Player A for his actions. I can learn about B's positive and negative reciprocity when looking at his rewards and punishments. Behaviour of Player A may carefully be interpreted as trust. The game is summarized in Figure 1. Player A and Player B are each endowed with 20 points at the beginning of the game. Player A has to decide whether to pass on 10 points to Player B, not to pass on any points, or take 5 points from Player B. The experimenter in any case doubles the transfer. Hence, if Player A passes on 10 points, Player B will receive 20 points. Conversely, if Player A takes 5 points, Player B will loose 10 points.

Figure 1: Experimental Design: Player A's choices vs. Player B's strategic options.



Examples: *- A takes 5, so B looses 10, then B punishes -10, he has to pay 10, so A looses 20.*
- A gives 10, so B gets 20, then B rewards +10, he has to pay 10, so A wins 20.

Player B can then decide how to respond to Player A's action. Player B can either reward or punish Player A. If Player B spends one point on rewarding Player A, Player A will receive two points. If Player B spends one point on punishing Player A, Player A will lose two points. Player B can use up to ten points for either punishing or rewarding Player A. I apply the strategy method for Player B, i.e., Player B makes a choice for each possible case that can arise (Player A passing on 10 points, passing on 0 points, or taking 5 points).

The sub-game perfect equilibrium with selfish preferences is easy to derive: Player B would never use any points to reward or punish, since Player A has already chosen his action. Consequently, Player A will maximize his payoff by taking 5 points from Player B (who looses 10 points). But behaviour is more interesting when agents have non-selfish preferences. Different social preferences can be examined for different players: First, the behaviour of Player B can be used to measure reciprocity. Player B can display positive reciprocity if he rewards transfers of 10 points relative to a neutral transfer (or if he rewards Player A for not taking any points away). Conversely Player B can display negative reciprocity if he punishes when Player A takes 5 points from him (or if he punishes Player A for not passing on any points).

Second, Player A's actions may be considered as trusting behaviour. However, they must be interpreted more carefully than in a standard trust game due to the ambiguous role of risk preferences. In a standard trust game (e.g. Glaeser et al., 2000; Fehr and List, 2004; Falk and Zehnder, 2006), players exhibit trust by transferring positive amounts to Player B. Risk aversion reduces transfers of first movers, because this reduces the variance in payoffs. In the moonlighting game Player A could also be motivated to pass on 10 points to Player B if he fears substantial punishment otherwise. Risk preferences have a more complicated impact on Player A's behaviour: A "neutral" action may be risky as well, as it may be rewarded or punished. These "outbursts" of reciprocity may create an even larger variance in the payment of Player A when he acts "neutral". Therefore the interpretation of Player A's actions has to be very careful and involve his beliefs concerning Player B's actions.

2.3.2 Treatments

Figure 2 summarizes the different treatments that have been conducted.

Student-Student Baseline: My control condition that is best comparable to other experiments is the Student-Student baseline treatment (referred to as S-S in the remainder of the chapter). 34 students participated as Players A, and 33 as Players B.[13] The students were told that their partners in the experiments were other students, not present now. Immediately after their choices, I elicited beliefs about the behaviour of their counterparts in the experiment, as I did in all treatments.

Student-Pilot (S-P): In this treatment, the choices of another 58 students as Players A were used. They were told that they would be matched with a pilot from the Swiss Air Force as

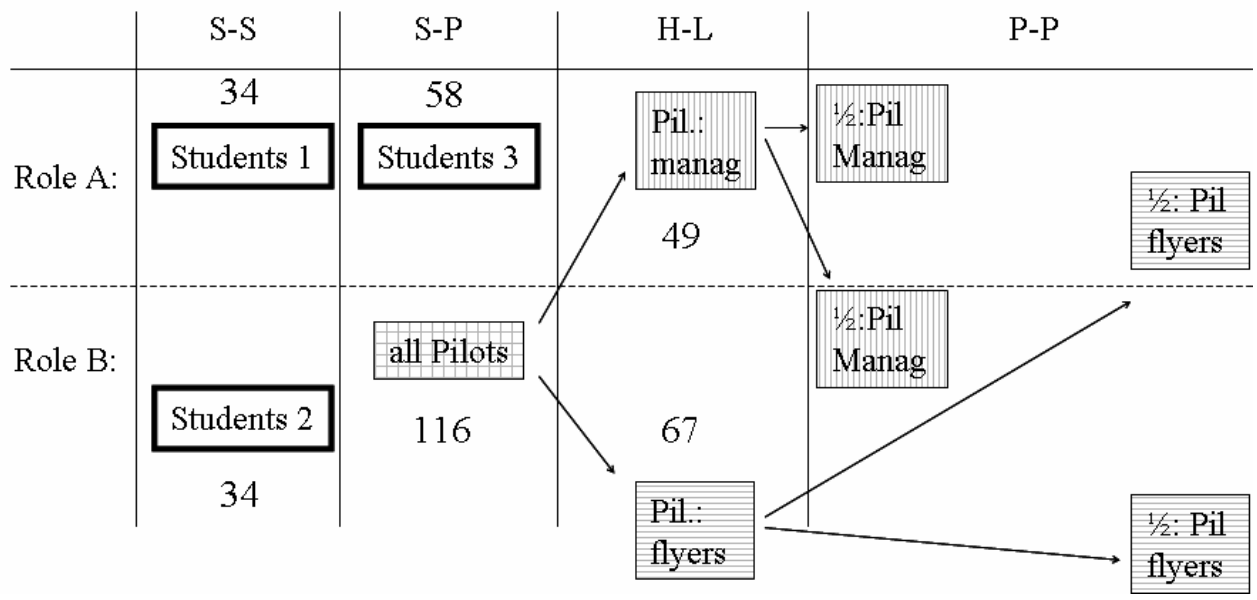
[13] The decision of the 34th player A was randomly assigned to the decision of a player B.

Player B. 112 pilots participated in this treatment. The students were informed that their choices may be used in multiple matchings, and that they would be paid the total amount they earned from all their matchings.[14]

Higher-ranking pilots - lower-ranking pilots (H-L): In this treatment, the pilots were assigned the roles of A players and B players based on their rank in the Air Force. Pilots with the rank of major or above (N = 49) were assigned the role of Player A and were matched with a pilot with the rank of captain or below (N = 67).[15] Some of the choices of Players A were used twice to determine all the payoffs of Players B.

Pilot-Pilot (P-P): Pilots in this treatments where paired with their peer-group. Lower-ranking pilots were paired with lower-ranking pilots, and told so. Similarly, higher-ranking pilots were paired with higher-ranking pilots and informed so.

Figure 2: Treatments and Subjects.



[14] As can be seen, a students' choice was used almost exactly twice.

[15] The separation between captains and majors is due to the fact that every squadron member becomes a captain after a certain amount of time, while squadron leaders and other superiors hold at least the rank of a major. Therefore the working level is represented by the maximum rank of a captain, while superiors are represented by the minimum rank of a major.

2.3.3 Procedures

The experiment for the baseline and the first part of the *S-P* treatment were conducted using students of the University of Zurich, who volunteered to participate as subjects of behavioural research studies. No economic students were allowed to participate in the sessions to avoid economic background bias due to their possible knowledge of game theory (Carter and Irons, 1991; Frank, Gilovich and Regan, 1993).

The data for the other treatments were collected on the occasion of a workshop for all Swiss Air Force pilots on the subject of "work satisfaction" and "burn out syndromes" on December 12, 2005. The workshop was compulsory, but a few flight missions could not be cancelled, and since the pilots cannot pick and choose whether they fly or attend the workshop, any participation bias seems unlikely.[16]

The pilots were seated according to their military rank. They were told that they had to make decisions and could earn money in doing so. Additionally they were informed that the data would be treated anonymously.[17] Thereafter the instructions and the decision sheets for the *S-P* treatment were distributed. Treatment *H-L* and *P-P* followed while every effort was made to ensure the subjects would not speak to each other and never hold more than one decision sheet at a time.

Only after the pilots had finished the last treatment were they issued a closed envelope with a copy of the decision sheet of a randomly assigned student to estimate their earnings from the *S-P* treatment. This late distribution of the partners' decision in the first treatment helped preventing a learning effect (Egas and Riedl, 2005). Payments where distributed at the end of the day.

2.3.4 Subject Characteristics

As former studies have shown, it is very important to obtain social background variables when comparing students to other subjects (Bellmare and Kroeger, 2005). Therefore the question of whether students can be compared to pilots in an efficient manner mainly depends

[16] This would rule out the argument of Lazear, Malmendier and Weber (2005) who found experimental results to be biased because people self-select into different settings.

[17] Special care was taken to ensure anonymity as not to endanger the actual team spirit of the pilots involved (Carpenter, Burks and Verhoogen, 2005). Furthermore, each subject was informed about his right to tell the experimenter not to use his decision-data for behavioural research.

on the overlap in the social background variables of the two populations. Table 1 shows the statistical means, minima, maxima, standard errors and numbers for pilots and students.

Table 1: Subjects characteristics: Pilots' and students' age, gender and education.

	Pilots					Students				
Variable	mean	min	max	sd	n	mean	min	max	sd	n
age	36.50	23	57	9.960	116	22.32	18	45	3.568	126
male	0.983	0	1	0.131	116	0.556	0	1	0.499	126
educ	12.71	12	19	1.552	116	12.5	12.5	12.5	0	126

Source: own calculations based on experimental evidence from December 2005.

The main difference between the two groups can be found in their mean age; the disparity is no less than 24 years. However, the two populations are overlapping, as the youngest pilot is 23 years old and the oldest student 45. Another difference can be found in the distribution of the genders within the two populations. Though there are only four female professional pilots in the Swiss Air Force, with only two of them participating in the workshop,[18] males were the more dominant gender among pilots than among students. As for education[19], the difference between students and pilots is small: On the one hand, I consider students to hold a high-school diploma, as this is an entry requirement to university.[20] Military pilot applicants, on the other hand, are currently required to hold a high school diploma or reach an equivalent educational level.[21]

Despite the above discrepancies in mean age and gender, I consider the two groups as sufficiently overlapping and therefore assume that statistical tests are not subject to out-of-sample predictions.

2.3.5 Behavioural Predictions

My design allows me to examine in detail how the behaviour of Player B (and to some extent of Player A) is affected by the different treatments.

[18] One female pilot was actually flying one of the missions that day and so could not attend the experiment. Another female pilot was working part-time for the Air Force and by coincidence was off duty on that particular day.

[19] Education was coded as theoretical years of education according the highest degree a subject has accomplished (see for example Bonjour (1997) or Falter and Ferro (2000)).

[20] No exact data on completed education was available for students.

[21] There are older pilots who have fewer years of schooling and a small group of pilots who have earned a university degree. These two groups almost offset each other.

The treatment most closely related to the previous literature is S-S: it provides me with the baseline measures of social preferences. Evidence from other studies suggest that when one controls for social background variables, students become comparable in their social preferences to the general public in their social preferences (Gueth, Schmidt and Sutter, 2002; Gachter, Herrmann and Thoni, 2004). Bellmare and Kroeger (2005) find that especially age, gender and education have significant effects on investments and rewards of a representative panel of the Dutch population and that after controlling for social background variables, no difference could be found between the representative panel and students.

Players B

Comparing the behaviour of Players B in treatment S-S to treatment P-P sheds light on my first issue, the preference for positive versus negative reciprocity. It shows the difference in social preferences between students and a group of subjects that interacts on a daily basis in an environment that yields high efficiency gains from cooperation.

There is strong evidence from previous studies that social ties between participants strongly influence behaviour towards more cooperation (see e.g. Burks, Carpenter and Goette, 2006).[22] Moreover, the scope for cooperation in the work environment of pilots may shape their social preferences. This has been suggested by Fehr and List (2004), who show more trustworthiness in managers of coffee planting cooperatives than in students.[23]

In light of possible negative consequences of negative reciprocity, I predict the following:

Cooperation among pilots is mainly based on stronger positive instead of stronger negative reciprocity. Therefore positive reciprocity is stronger and negative reciprocity is weaker among pilots than among students.

My second focus lies on hierarchy: I compare Player B's behaviour in P-P to H-L treatment, where I match low-ranking pilots with high ranking-pilots.

It can be argued that ranks between pilots matter, as only pilots of similar rank are similar in many dimensions, including age and years of service. Further, similar ranking pilots are more likely to interact with each other while performing similar tasks, since the higher-ranking

[22] Burks, Carpenter and Goette (2006) find that bike messengers cooperate more as second movers in a sequential prisoners' dilemma even to the extend that 30 percent cooperate in spite of defection of the first-mover.

[23] Holding transfers of the first mover constant, the managers in their experiment return significantly more than the student control group. In contrast to the standard trust game that has been applied by Fehr and List (2004), my experiment lets me distinguish between positive and negative reciprocity.

pilots mostly pursue management duties within the Air Force. Thus, I might expect a higher social distance when pilots of "unequal" rank are matched in an experiment. Further evidence from Hoff and Pandey (2006) and La Porta et al. (1997) show that social hierarchy can have a strong impact on behaviour.

However, as perceived hierarchy might bear negative externalities in such an environment as discussed in the introduction of this chapter, I predict the following:

Social distance induced by hierarchy does not affect positive and negative reciprocity of pilots.

To finally test whether high levels of cooperation are based on group-specific norms, I increase the social distance to a degree of total outsiders. The comparison of P-P treatment vs. S-P treatment shows whether pilots' degrees of reciprocity are different when interacting with a student rather than with a pilot. There is strong evidence from earlier studies that pro-social behaviour is different when individuals interact with someone they perceive to belong to a different social group. Most closely related, Goette, Huffmann and Meier (2006) find evidence consistent with in-group favouritism when officer candidates in the Swiss Army interact with a member of their own platoon relative to when they interact with a member of a different platoon.[24] Similarly Ruffle and Sosis (2006) find that members of Kibbutzim are more cooperative in interactions with member of Kibbutz than with subjects living in "normal" cities.[25] Fershtman and Gneezy (2001) finally find that subjects' social preferences strongly depend on demographic differences to partners in experiments. Hence I predict the following:

When pilots are paired with students positive reciprocity decreases and negative reciprocity increases compared to pilots interacting with fellow pilots or superior pilots.

Player A

In the case of Player A, I am not only interested in his actions. Beliefs might be important as well, as they influence the actions of Player A.[26] Trusting behaviour can be explained as follows: If Player A believes in a strong positive reciprocity from Player B, he is prone to

[24] Bernhard, Fehr and Fischbacher (2006) find similar results for self-selected groups: Members of clans in Papua-New Guinea show the same behavioural tendencies when interacting with somebody from their own clan relative to interactions with somebody from a different clan.

[25] Further, when Kibbutz members interact with city members, cooperation is at the same level as that of city members.

[26] This is due to the fact that for player A, the strategy method could not be applied

express more trust by transferring more money to Player B.[27] Additionally, in the moonlighting game he might be influenced by the fear of negative reciprocity of Player B. If, however, his transfer is high in absence of corresponding expectations about Player B's back transfer, his action might be motivated by risk preferences. These are ambiguous as discussed in section 2.3.1.. The fact that military pilots have a risk of death, which is at least 60 times higher than their student counterparts, could play a role in this case.[28] Hence, only if believed positive reciprocity is high, a statement in the favour of trust can be made. But even then caution has to be exercised when interpreting the results because risk preferences might be interacting with the beliefs of Player A.

The comparison of S-S vs. P-P is used to establish the differences in stated beliefs and actions between pilots and students.[29] In pursuing the argument for Player B I propose the following:

Pilots exhibit more trust: They anticipate higher levels of positive reciprocity, believe in higher back transfers and thus transfer more than students.

As a final point I test the effect of in-group favouritism by introducing hierarchy when comparing treatment P-P to H-L. One might argue that social distance hinders trust (Hoffmann, McCabe and Smith, 1996). However, as cooperation across ranks is essential among pilots in the same crew to remain effective, social distance should not be critical for the actions of pilots. Hence I predict the following:[30]

Beliefs and transfers of Players A remain the same whether pilots are paired with lower ranking pilots or with pilots of the same rank.

[27] Fehr and List (2004) find higher transfers of first movers in their trust game with managers. In Burks, Carpenter and Goette (2006) bike messengers cooperate more as first movers than students in a sequential prisoners dilemma.

[28] The mean risk of death for a Swiss worker in 2001 amounts to 0.5 deaths per 10'000 workers. In civil aviation it amounts to 1.5 deaths per 10'000 workers (SUVA, 2004). For Swiss Air Force pilots, it amounts to 33 deaths per 10'000 pilots each year for the period of 1981 until 2001. During the period of 1941 until 2001 the Swiss Air Force had even lost 81.5 pilots to crashes per 10'000 pilots each year. As some of the older pilots among my subjects may have based their decision on early statistics, even this number could be claimed representative.

[29] Although I do not have a treatment that introduces experimental variation in beliefs to identify the beliefs' effects on behaviour the comparison of pilots and students beliefs can be telling.

[30] As I do not have a P-S treatment (see Figure 2 on page 11), I cannot test for total strangeness for player A.

2.4. Results

2.4.1 Comparison of S-S to Previous Literature

The S-S treatment serves as a benchmark to previous studies with students as subjects. In my experiment, the B player punishes negative transfers with 3.6 points, rewards zero transfers with essentially zero (0.3) points, and rewards a transfer of 10 points with 5.6 points on average. In relative terms, my subjects thus invested 36% of the points they lost into punishment and 28% of the points they gained into rewards. The latter figure is strikingly close to Abbink, Irlenbusch and Renner (2000), who find a corresponding figure of 26%. Evidence in Abbink, Irlenbusch and Renner (2000) points to a slightly stronger negative reciprocity than my study, although this could be related to the details of the game.[31]

In my base treatment, 14% of players B choose to take 5 points, 27% choose to transfer 0 points and 59% of players transferred 10 points. In Abbink, Irlenbusch and Renner (2000) 18.8 % choose a negative amount (12.5% more than half of the possible points), 15.6% choose to send 0 points and 65.6% sent a positive amount (43.8% more than half of the possible points).[32]

2.4.2 Behaviour of Player B

- ***Result 1:*** *Pilots are significantly more positive reciprocal and slightly less negative reciprocal towards pilots than students towards students.*

Evidence for this result can be found in Figure 3, which shows the average back transfers to different actions of players A. A clear difference between back transfers of pilots to back transfers of students can be found. Particularly, kind actions of players A are more rewarded

[31] The version of the moonlighting game in Abbink, Irlenbusch and Renner (2000) give the A players a continuous choice to either transfer or extract points from B, hence only a comparison of the relative shares is possible. The continuous nature of the game also makes it possible to extract only a few points from player B, which usually are punished quite severely. This may explain why there is more evidence for negative reciprocity in Abbink, Irlenbusch and Renner (2002). Furthermore for player A positive transfers are tripled, while negative transfers are not altered by the experimenter. In the case of player B, only negative transfers are tripled. The moonlighting game of Cox, Sadiraj and Sadiraj (2002) shows further differences: Players B can not only reward or punish but even take money from player A. Cox, Sadiraj and Sadiraj (2002) find that 38% of their players B, who have lost money, take revenge on player A. Players B, who have made money, reinvested 39% of the money they have gained in to rewards.

[32] In Cox, Sadiraj and Sadiraj (2002) 43% choose a negative amount, 10% send 0 points and 47% a positive amount.

by pilots than students, which points to stronger positive reciprocity. Additionally, unkind actions of players A are less revenged by pilots than by students, which points to weaker negative reciprocity.

Figure 3: B's Actions: S-S vs. P-P treatment.

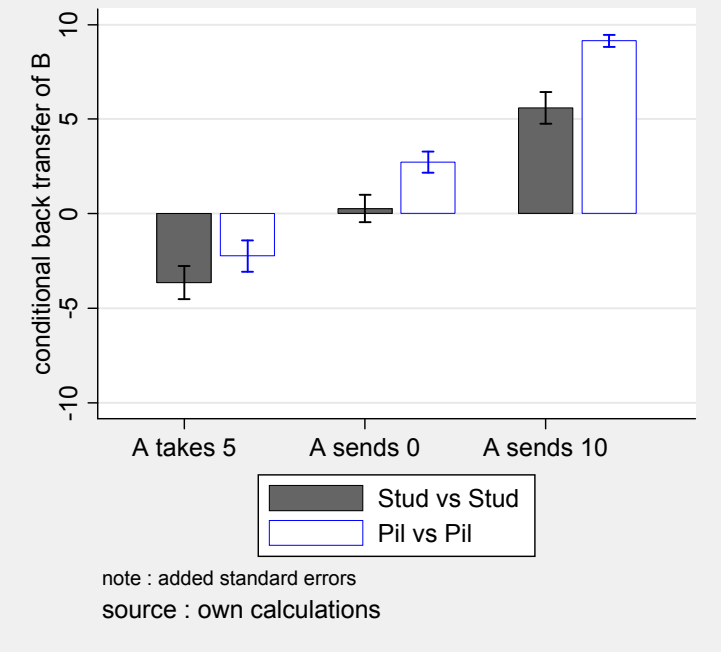


Table 2 presents a more stringent statistical test.[33] An OLS regression is performed with heteroscedastic standard errors. The regression controls for age, squared age, gender and possible differences in education. According to the results in columns 1 through 3 of Table 2, variation in gender, age or education between students and pilots do not explain the whole difference in back transfers. Furthermore, it cannot be rejected ($p < 0.01$) that pilots show a more positive reciprocal nature than students. For negative reciprocity on the other hand I do not find a statistically significant difference between pilots and students. I therefore infer that the fact that negative reciprocity of pilots is not stronger compared to those of students cannot be rejected. To further substantiate my findings I added a conservative significance test,

[33] As a paper and pencil method was applied, I am not able to prevent missing actions and beliefs. In addition to that, some individuals stated beliefs that added up to more than 100% and where therefore disregarded. Finally, few individuals probably misinterpreted the decision sheets. They stated decisions that are exactly the opposite of reciprocal, as they rewarded unfriendly and punished friendly actions. I believe that I should not consider these data for my analysis. However, additional regression results show that none of my findings are sensitive to the inclusion of these data. Exact tables of missing and used observations and beliefs can be found in Table 16 through Table 19 in the appendix.

which adjusts for the fact that I test multiple hypotheses (Holm, 1979, as cited in Romano and Wolf, 2005). The data are robust to this test, as the significance level for positive reciprocity is not affected.[34]

Table 2: B's actions conditional different actions of A
(Comparison of S-S and P-P treatments)

	B's action conditional a negative action of A	B's action conditional a neutral action of A	B's action conditional a positive action of A
pilot	1.996	2.516	3.974$^{###}$
	(2.212)	(1.719)	(1.081)***
male	-1.023	0.804	3.132
	(1.643)	(1.414)	(1.428)**
age	-0.476	-0.117	-0.555
	(0.508)	(0.348)	(0.265)**
age^2	-0.007	0.001	0.006
	(0.007)	(0.005)	(0.003)*
educ	-0.243	0.246	0.162
	(0.393)	(0.379)	(0.151)
const.	-6.928	-1.147	11.364
	(9.798)	(7.277)	(4.488)**
R^2	0.056	0.092	0.333
Prob > F	0.464	0.132	0.000
n	88	88	88

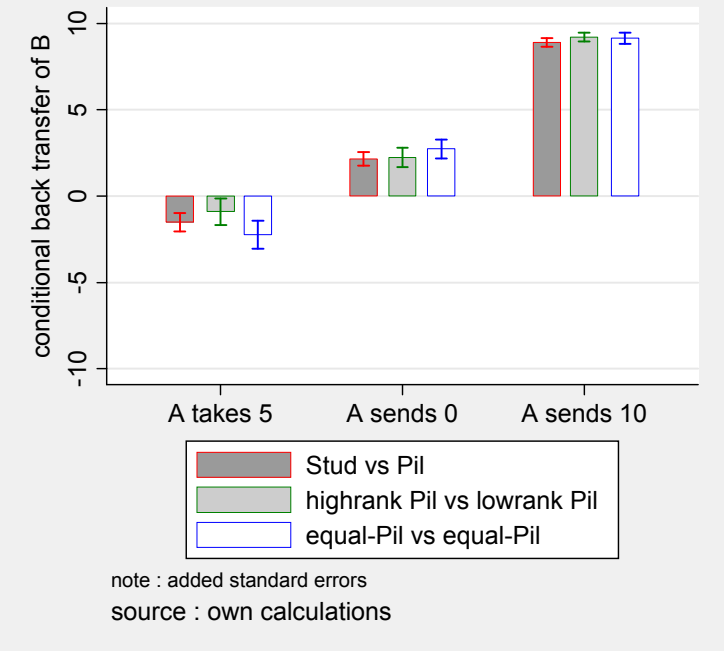
***Notes*: Dependent variable: B's transfer after different actions of A. Coefficients of OLS-regression**
(Robust standard errors in parentheses).
***Level of significance:* *0.1<p, **0.01<p<0.05, ***p<0.01,**
corrected levels for multiple hypothesis (Holm, 1979):
$^{#}$0.1<p, $^{##}$0.01<p<0.05, $^{###}$p<0.01
***Source:* own calculations based on experimental evidence from December 2005.**

The data also show some demographic differences. Age appears to have an effect on positive reciprocity, which follows a u-shape with a minimum at 46 years. These findings are very

[34] To further assess the difference in social preferences between pilots and students, I examine the costs the subjects are willing to incur to respond to players A. The OLS regression results with the dependent variable being the absolute amount of money spent are represented in columns 1 through 3 of Table 15 in the appendix. The only significant difference I find between pilots and students is in the response to a positive transfer of 10 points. Comparing column 1 of Table 15 to Table 2, I might infer from the fact that the point estimate drops from 2 to 0 points that pilots are not generally less engaged in interactive behaviour, but do it in a less negative way when treated unfriendly. This result gives some support to the assumption that pilots may be less negatively reciprocal.

close to those of Bellmare and Kroeger (2005). This is important because they use a representative panel of the population. In their trust game, they find a similar relation between age and the propensity to reward investments with a minimum between 35 and 40 years of age. A further similarity in results exists for the dummy variable "male", which has a positive effect on back transfer when treated kindly (column 3).[35] Lastly, education has no statistically significant effect on back transfers in my results, which might be caused by the low variation in my data. Bellmare and Kroeger (2005), however, find a negative educational effect for responders in an investment game. Due to the otherwise high comparability of my results to Bellmare and Kroeger (2005), who do not rely on student data only, I argue that the behaviour of my subjects might be in fact special compared to social preferences among the general population in an anonymous setting.

Figure 4: B's Actions: S-P vs. H-L vs. P-P Treatment.



To examine the exact nature of the cooperation differential I introduce social distance, first in the form of hierarchy by pairing high-ranking pilots with low-ranking pilots and finally in the form of total strangeness by pairing students with pilots.

[35] Equal results can be found in Gueth, Schmidt and Sutter (2002).

- ***Result 2****: Cooperative behaviour of pilots, mainly based on strong positive reciprocity, is not confined to fellow pilots. There is no difference in the back transfers of pilots to higher-ranking pilots or even students.*

Figure 4 shows virtually no difference between the back transfers of treatments P-P, H-L and S-P.[36] More stringent results are displayed in Table 3 where OLS-regressions with cluster analysis have been run. Dummy-variables have been added for the S-P and the H-L treatment, the P-P treatment being the base group in this case.

Table 3: B's actions conditional different actions of A
(Comparison of P-P, S-P and H-L treatments)

	B's action conditional a **negative** action of A	B's action conditional a **neutral** action of A	B's action conditional a **positive** action of A
S-P	0.725	-0.482	-0.227
	(0.719)	(0.525)	(0.327)
H-L	0.284	-0.336	-0.073
	(0.386)	(0.253)	(0.148)
≥major	0.593	-0.394	-0.552
	(2.075)	(1.578)	(0.874)
male	-2.505	0.447	0.709
	(3.255)	(2.160)	(0.336)**
age	0.229	0.510	-0.007
	(0.457)	(0.321)	(0.167)
age^2	-0.005	-0.007	-0.000
	(0.006)	(0.004)*	(0.002)
educ	-0.069	0.328	0.231
	(0.289)	(0.288)	(0.088)***
const.	-0.544	-10.476	6.033
	(9.766)	(7.330)	(2.974)**
R^2	0.052	0.045	0.033
Prob > F	0.225	0.286	0.073
n	233	233	233
clusters	113	113	113
Prob >F for S-P and H-L	0.587	0.408	0.784

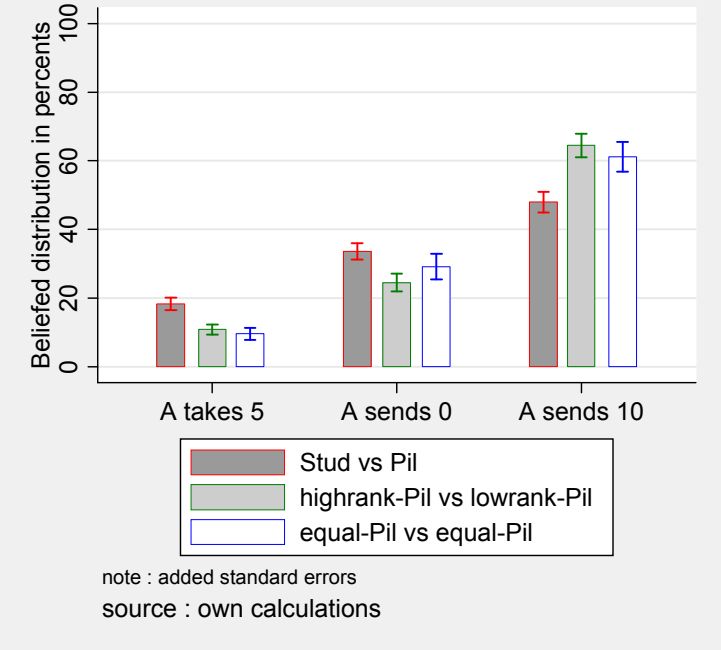
***Notes*: Dependent variable: B's transfer after different actions of A. Coefficients of OLS-regression (Robust standard errors adjusted for clustering on individuals in parentheses).**
***Level of significance:* *0.1<p, **0.01<p<0.05, ***p<0.01, corrected levels for multiple hypothesis (Holm, 1979): #0.1<p, ##0.01<p<0.05, ###p<0.01**
***Source:* own calculations based on experimental evidence from December 2005.**

[36] There is no concern for effects of subjects playing both roles in the experiment (see for example Burks, Carpenter and Verhoogen, 2003), as this is not the case in the same treatment. The procedure, however, does not exclude a possible order effect, which is minimized by not informing the subjects about their earnings before the end of the last treatment.

According to columns 1-3, I must conclude that the two treatment effects are not statistically significant. No differences between the three treatments where detected. The precision of this result is fairly high, as differences above 0.6 points would have been picked up. The respective F-tests confirm the result.[37]

It is interesting to ask whether pilots perceive any difference at all between pilots and students playing part A.[38] According to Figure 5 pilots have identical beliefs about pilots of the same or higher rank. By contrast, their beliefs about students' behaviour are more negative. They expect a lower percentage of the students to transfer 10 points, and a higher percentage to take 5 points.

Figure 5: B's Beliefs about A's Actions in different Treatments.



[37] To detect if superiors behave differently from normal pilots, probably due to higher salaries, I introduced an additional variable for ranks starting with major. As no statistical significant difference can be detected I furthermore conclude that there is also no wealth effect (as majors earn more than captains) involved in the equation, which confirm findings of Slonim and Roth (1998) and Bellmare and Kroeger (2005).

[38] This has been suggested by Manski (2002) and applied by Bellmare, Kroeger and van Soest (2005). However, as in my case the strategy method has been applied I do not consider using beliefs as an independent variable in the regression. Every pilot had the chance to make conditional decisions, which cannot depend on his beliefs of the distribution of the actions of first movers. Hence the actual decision of player A cannot in any logical sense influence the decision of player B.

Again, Table 4 presents statistical tests. OLS-regressions with cluster analysis confirm the results. This shows that pilots are well aware of their partners and all the same do not change their behaviour.

Table 4: B's beliefs about percentages of players A making different actions. (Comparison of P-P, S-P and H-L treatments)

	B's Beliefs about % of A making a **negative** transfer	B's Beliefs about % of A not transferring anything	B's Beliefs about % of A making a **positive** transfer
S-P	8.584$^{###}$	4.729	-13.314$^{###}$
	(1.922)***	(3.162)	(3.827)***
H-L	0.606	-0.038	-0.569
	(0.904)	(1.881)	(2.267)
≥major	-8.259	1.670	6.589
	(4.226)*	(8.935)	(10.239)
male	-5.247	10.530	-5.283
	(2.591)**	(3.045)***	(4.048)
age	-0.151	3.577	-3.426
	(1.261)*	(1.609)**	(2.098)
age^2	0.009	-0.039	0.030
	(0.017)*	(0.021)*	(0.028)
educ	0.771	0.112	-0.883
	(1.225)	(1.431)	(1.908)
const.	1.660	-57.558	155.898
	(21.575)	(32.947)*	(44.080)***
R^2	0.089	0.110	0.149
Prob > F	0.000	0.000	0.000
n	204	204	204
clusters	101	101	101

***Notes*: Dependent variable: B's belief about how many percents of A making different actions. Coefficients of OLS-regression (Robust standard errors adjusted for clustering on individuals in parentheses).**
***Level of significance:* *0.1<p, **0.01<p<0.05, ***p<0.01**
corrected levels for multiple hypothesis (Holm, 1979):
$^{#}$0.1<p, $^{##}$0.01<p<0.05, $^{###}$p<0.01
***Source:* own calculations based on experimental evidence from December 2005.**

It is tentative to conclude that the Swiss Air Force is able to screen, train or socialize its pilots in a way that cooperation is primarily sustained through higher levels of positive reciprocity, and the effect is robust to social hierarchy. It is astonishing, however, that pilots do not change their behaviour when paired with outsiders. This result is surprising, as other studies have found strong effects of social distance (La Porta et al., 1997; Fershtman and Gneezy, 2001; Ruffle and Sosis, 2006; Hoff and Pandey, 2006; Bernhard, Fehr and Fischbacher, 2006; Goette, Huffmann and Meier, 2006). Therefore I conclude that differences in the behaviour of

pilots compared to other subjects are not mainly caused by social ties but rather by more cooperative preferences of pilots.

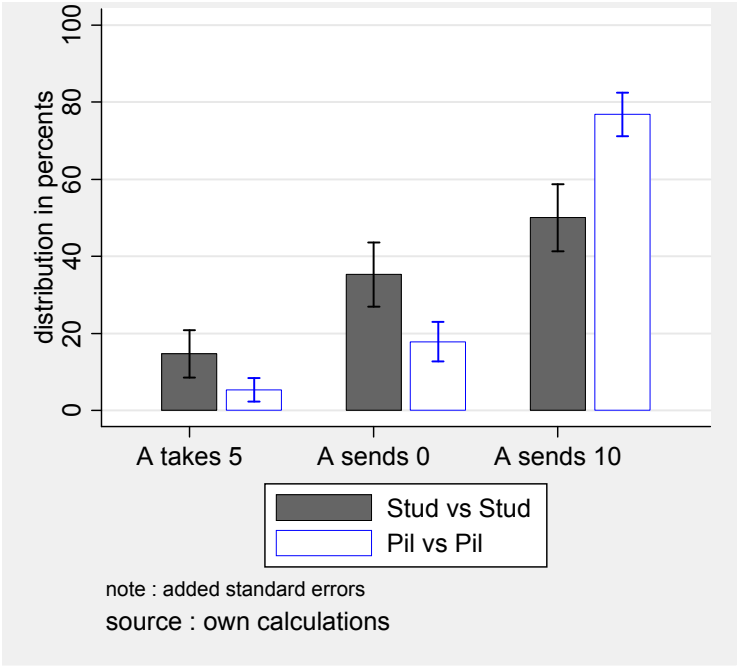
Even though the situation is less clear for Player A, the next subsection analyses his behaviour.

2.4.3 Behaviour of Player A

- ***Result 3:*** *Pilots do not seem to transfer more than students.*

The formal statistical test can be found in the second column of Table 5,[39] where the amount transferred to Player B is regressed on the treatment condition. I do not find a statistically significant difference between the transfer of pilots compared to students when controlling for gender, age, age^2 and education. The maximum-likelihood ordered probit estimation in the next column of Table 5, which accounts for the restricted choice characteristic for Player A, confirms the results.

Figure 6: A's Actions: S-S vs. P-P Treatment.



If I compare these results to Figure 6, which shows the different distributions in percents of choices of students and pilots in the position of Player A, I am astonished: As I had originally

[39] The 4 observations that had no respective beliefs (see Table 16 and Table 19 in the appendix) had to be dropped to keep estimations comparable. This caused no significance difference in the results.

Table 5: A's actions / A's beliefs about B's back-transfer conditional different transfers of player A / A's actions according beliefs (Comparison of P-P and S-S treatments)

	A's action			A's action according Belief		A's Belief about B's back transfer					
	OLS		ordered Probit	OLS	ordered Probit	after a **negative** transfer		after a n**eutral** transfer		after a **positive** transfer	
P-P	3.467	1.298	0.293	-0.706	-0.218	2.550	1.722	1.975	2.008	3.891##	3.165
	(1.655)**	(1.896)	(0.445)	(1.875)	(0.476)	(1.539)	(1.743)	(1.032)*	(1.277)	(1.356)***	(1.847)*
male	0.796	0.977	0.228	1.501	0.473	0.784	0.853	-0.139	-0.142	-0.702	-0.641
	(2.223)	(2.110)	(0.402)	(2.027)	(0.419)	(1.334)	(1.336)	(0.714)	(0.733)	(1.456)	(1.487)
age	-0.053	0.809	0.183	0.777	0.222	-0.048	0.281	-0.033	-0.047	-0.082	0.206
	(0.065)	(0.494)	(0.124)	(0.499)	(0.128)*	(0.058)	(0.410)	(0.042)	(0.273)	(0.038)**	(0.344)
age^2		-0.011	-0.003	-0.010	-0.003		-0.004		0.000		-0.004
		(0.065)*	(0.002)	(0.007)	(0.002)*		(0.005)		(0.003)		(0.004)
educ	-0.006	0.021	0.027	-0.119	-0.039	-0.510	-0.499	-0.187	-0.187	0.239	0.248
	(0.500)	(0.383)	(0.089)	(0.520)	(0.117)	(0.338)	(0.299)*	(0.153)	(0.151)	(0.234)	(0.267)
A's belief after pos action				0.490	0.093						
				(0.264)*	(0.055)*						
A's belief after neut action				0.382	0.131						
				(0.263)	(0.077)*						
A's belief after neg action				-0.182	-0.062						
				(0.124)	(0.034)*						
cut1			3.131		3.962						
			(2.239)		(2.435)						
cut2			2.230		2.916						
			(2.206)		(2.432)						
const.	5.244	-8.475		-10.514		2.261	-2.972	3.039	3.252	5.109	0.521
	(7.091)	(9.119)		(9.833)		(4.427)	(7.060)	(1.797)*	(4.073)	(3.042)*	(6.700)*
(pseudo) R^2	0.076	0.106	0.063	0.266	0.171	0.068	0.073	0.055	0.055	0.143	0.150
Prob > F	0.180	0.158		0.010		0.115	0.163	0.173	0.232	0.010	0.004
Prob > Chi^2			0.219		0.045						
n	86	86	86	86	86	86	86	86	86	86	86
Prob >F				0.082							
Prob > Chi^2 for all Beliefs					0.048						

***Notes*: Dependent variable: First Column: A's action (-5, 0, 10), coefficients of OLS-regression. Second column: A's action, coefficient of ordered Probit-regression, third through fifth column: A's belief about B's back-transfer conditional different actions of A. Coefficients of OLS-regressions. Sixth column: A's transfer to B, coefficients of OLS-regression, seventh column: coefficient of ordered Probit-regression (Robust standard errors in parentheses).**

***Level of significance:* *0.1<p, **0.01<p<0.05, ***p<0.01 corr. levels for mult. hypo. (Holm, 1979): #0.1<p, ##0.01<p<0.05, ###p<0.01**

***Source:* own calculations based on experimental evidence from December 2005.**

expected, clearly more pilots than students send 10 points to Player B. But the control variables of the regression seem to explain the whole difference.[40] Furthermore the estimators for age point towards an inverted u-shape of the age effect with a maximum at the age of 37. This result is very comparable to those of Bellmare and Kroeger (2005), who find in their representative panel an inverted u shaped age-effect with a maximum at 37 years as well.[41] Hence I conclude that pilots do not seem to transfer more than students as players A, as any difference is already explained by the control variables.

As I expected that pilots transfer more because they believe in higher back-transfers of fellow pilots, it is interesting to study whether actions of players A correlate with different beliefs about the conditional back transfer of players B (Camerer and Fehr, 2006; Goette, Huffmann and Meier, 2006).[42]

- ***Result 4****: Actions seem to be mainly motivated by players' beliefs in positive and negative reciprocity, i.e. by more optimistic beliefs about back-transfers.*

An OLS and ordered probit regression incorporating subjects' beliefs as independent variables are displayed in the fourth and fifth column of Table 5. They show that indeed beliefs are statistically significant factors for subjects' actions.[43] All the more, the r-squared climbs from 11% (in the regression without beliefs) to 27%, which indicates the high relevance of the beliefs. Finally, the relevant combined F-tests confirm the results. It appears as if higher transfers are motivated by more optimistic beliefs about positive, neutral and negative actions.[44]

---

[40] To find out, whether there is a good reason to include the control variables, I run different regressions and found that finally the squared age effect is responsible for the high p-value of the treatment factor. To show the effect, I included in the first column of Table 5 a regression without this factor, where the statistical effect of the P-P treatment is quantitatively large and significant as originally expected. However, as age^2 has a p-value of 0.088 and age one of 0.105, I may not exclude the factor.

[41] Additionally Alesina and La Ferrara (2002) find similar effects. A possible explanation can be found in Glaeser, Laibson and Sacerdote (2002), who argue that investment in social capital follows the same pattern as investment in human capital.

[42] Fehr et al. (2006) accordingly find in a trust game that not all the differences in transfers between Americans and Germans can be explained by different beliefs.

[43] The respective p-values for the factors of the two regressions are not very different but climb over the 10% margin for the OLS-regression in the case of A's beliefs after neutral and negative actions.

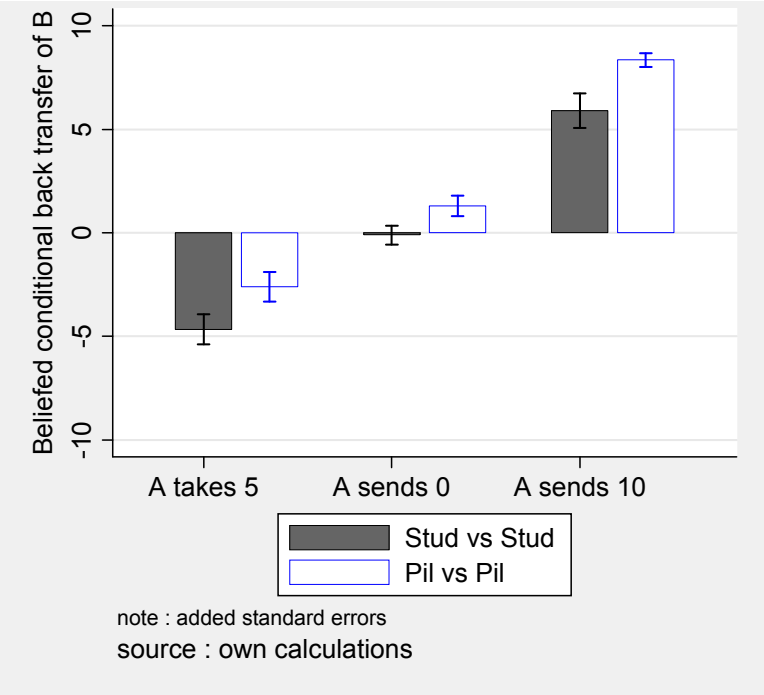
[44] However, one has to be cautious interpreting these results as beliefs might be affected by risk aversion. Furthermore as their beliefs have been asked after they had made their decision, it seems possible that the subjects acted according their intuition and when asked about their expectations, justified their decisions through

Finally I study the beliefs of players A.

- ***Result 5****: There is weak evidence that Pilots believe in higher back-transfers than students.*

The formal statistical test can be found in the last columns of Table 5, where the beliefs of players A about back-transfers of Player B after negative, neutral and positive transfer are regressed on the treatment condition. I find a statistically significant difference between the beliefs of pilots and students when controlling for gender, age, and education. The estimated difference in believed back-transfers after a positive action is 3.9 points and is statistically significant with a p-value of 0.005.[45] However, as I include the squared effect of age, only beliefs about back-transfers after positive actions are positive and marginally statistically significant with a p-value of 0.090.[46] This result is astonishing when compared to Figure 7 where clearly pilots believe in higher back-transfers of Players B than do the students.

Figure 7: A's Beliefs: S-S vs. P-P Treatment.



It seems unclear whether pilots do not transfer more because they do not have more optimistic beliefs about back-transfers than students, or whether the ambiguous effects of risk preferences in the moonlighting game offset the slightly more optimistic beliefs of pilots.

---

appropriate beliefs (Kahnemann, 2003), although the experimenter promised the subjects money for accuracy in beliefs.

[45] When I apply the multiple hypotheses test (Holm, 1979) the p-value climbs to 0.015.

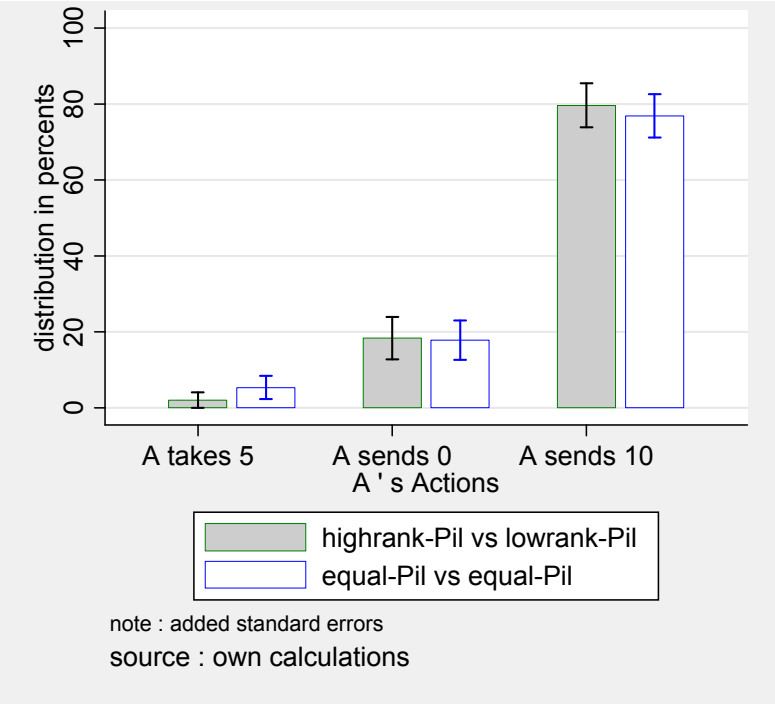
[46] They even become statistically insignificant if the multiple hypothesis-test is applied (Holm, 1979).

To finally test for social distance due to hierarchical effects, I compare the P-P to the H-L treatment.

- ***Result 6:*** *Higher-ranking pilots do not trust less in lower-ranking pilots and therefore do not transfer less to lower ranks, compared to pilots interacting with equal ranks.*

Only graphical evidence for this result can be given in Figure 8 as just four players A changed their action from H-L to P-P treatment. Surprisingly it was all to the worse of their partners. The results complement the previous findings from players B. They suggest that hierarchy in this setting has no disruptive effect on social preferences. This must be one feature of the very special work environment even though it is within the armed forces where hierarchy plays a more influential role than in the civilian world (Akerlof and Kranton, 2005).

Figure 8: A's Actions: H-L vs. P-P Treatment.



2.5. Conclusions

Cooperation within its workforce is a highly important factor for almost every firm in the market (La Porta et al., 1997). However, it is difficult to enforce cooperation among workers: It is hard to observe and therefore unlikely to be enforced by incentives created by repeated interactions between workers and firms. Hence high levels of cooperation must be established on account of higher grades of positive or negative reciprocity.

Previous studies have shown that negative reciprocity is a strong behavioural force when subjects are mostly college students (Charness and Rabin, 2002). However, negative reciprocity among workers may bear inefficient outcomes for a firm. Furthermore, if workers of different levels have to cooperate, perceived hierarchy may hinder cooperation. Therefore I hypothesized that firms could assure cooperation by appropriate screening, training or socialization of its workers.

To test this hypothesis, I use a subject pool of professional Swiss Air Force pilots, who must arguably cooperate very closely to achieve their mission goals. To detect both positive and negative reciprocity I apply the moonlighting game (Abbink, Irlenbusch and Renner, 2000).

I find that pilots are indeed significantly more positive reciprocal and slightly less negative reciprocal towards pilots than students towards students.[47] Furthermore the estimators of my control variables are very comparable to those of Bellmare and Kroeger (2005), who use data of a representative panel of the Dutch population. Therefore I infer that my results are of general relevance. I can furthermore show that trust may depend on the level of expected reciprocity. However, the higher levels of trust expressed by pilots seem to be mostly explained by control variables. But this unexpected result may be biased due to the ambiguous effects of risk preferences in the moonlighting game.

To test for the effects of perceived hierarchy, I conduct further experiments and find that pilots' stronger positive and weaker negative reciprocity is not confined to the interaction with fellow pilots. Pilots do not treat higher-ranking pilots different than fellow pilots.

I finally study how pilots behave towards strangers. I find that pilots do not change their behaviour, as there is no difference between their back transfers to students and their back transfers to their fellow pilots. Thus my results suggest that these pilots have different preferences and are even outside their environment more cooperative. This is astonishing in light of previous studies that indicate strong effects of group identity (Gachter and Fehr, 1999; La Porta et al., 1997; Fershtman and Gneezy, 2001; Ruffle and Sosis, 2006; Hoff and Pandey, 2006; Bernhard, Fehr and Fischbacher, 2006; Goette, Huffmann and Meier, 2006).

These findings indicate that the Swiss Air Force is in fact able to screen, train or socialize its pilots in a way that cooperation among its workers is primarily maintained through positive

[47] This Result does not contradict Mas (2006) who finds strong negative reciprocity among police agents. It seems impossible in his set-up to detect positive reciprocity at all, as either the expectations of the policemen are met or not met but never exceeded. Furthermore he researches the interaction of firms and workers, while in my case I examine the interaction between workers and workers.

reciprocity, and hierarchy does not compromise teamwork. Future research should address this issue and try to separate these different possible channels.

Not surprisingly there is evidence which points towards the screening thesis: In Hedinger (2004) I showed that reserve pilots of the Swiss Air Force that work in non-flying jobs (e.g. as engineers or medical doctors) and pass the same screening as the professional pilots earn significantly more than comparable individuals with an equal job.[48] This could be interpreted as an exceptionally high level of productivity, which may be based on high teamwork abilities. Hence, if reserve pilots show the same levels of reciprocity, cooperation could be claimed to screening effects.

The findings of this chapter imply that an employer who claims having such a workforce does not need to enforce cooperation among its workers. Furthermore, he might even profit from the exceptionally high levels of positive reciprocity, as his employees do not distinguish between superiors or fellow workers. To find out whether this is true I focus in the next chapter on the applicability of reciprocity to the relations of employees and employers. I use the example of subjective performance evaluations.

[48] After including various control variables, propensity-matching estimates a salary difference of 25%.

CHAPTER 3:

Reciprocity Effects of Subjective Performance Evaluations

A Case Study with personnel records

3.1. Introduction

For managers, the ability to predict the possible consequences of the implementation of different incentive systems is very important. In the case of the direct incentive system, it has long been known that this system only enhances the performance that is measured (Gibbons, 1998: 115f; Prendergast, 1999:8). Consequently, the more comprehensive *subjective performance evaluation* system has been suggested. This should induce the optimal allocation of time and effort to the multiplicity of tasks in most of today's jobs. But this system has rarely been researched by economists (Prendergast, 1999:9;11;22;56-57). Data and measurable factors seem to be very difficult to obtain, perhaps also because evaluations are of a subjective nature. With this chapter I will try to close part of this gap by making use of an unusual situation: I use a dataset of subjective performance evaluations in a "multi-task"[49] environment, combining it with operational data that has not been available to the evaluating persons. They were therefore unable to directly observe specific tasks. Hence, this unobserved effort might not affect performance evaluations. This situation renders me the perfect opportunity to find out whether subjective performance evaluations have spillover effects on

[49] I use the term "Multi-Task" according to Prendergast (1999:22), who does not mean that a worker has to perform many tasks at the same time, but that he has to choose to which of many tasks he primarily allocates his time and effort.

the provision of unobserved effort in the next period. To predict these effects, I use the insights of a growing literature in reciprocity that is mainly based on experiments. It may help to shed some new light on how a subjective incentive system influences the provision of unobserved effort.

I believe the subject is relevant for more than just this specific situation because most systems using subjective performance evaluations may not be able to consider all aspects of performance or effort. Today's multi-tasking environments are just too complex. Standard incentive system theory would predict that unaccounted effort would vanish because it does not enhance a worker's salary (Prendergast, 1999:22-29). But in a system with subjective performance evaluations, a worker does not solely depend on his shown performance alone, but is also, to some extent, at the mercy of his superiors if not all tasks are fully verifiable. This fact may give rise to considerations of fairness and reciprocity. A worker might react positively to a good evaluation or negatively to a bad one. As the worker in a long-term contract does not want to punish himself by endangering his subjective performance evaluation, he would most probably reward or retaliate in the field of unobserved effort. I consequently researched whether employees tend to react reciprocally by means of unobserved effort to changes in their subjective performance evaluations and if so unobserved effort may be sustainable.

To research this question I use a combined longitudinal dataset of subjective performance evaluations with the recorded working hours of all Swiss Air Force pilots. I received the data from my former direct superior, the commander of these aviators.[50] The specific situation of the Swiss Air Force pilots is the following: They not only have to fly a lot, but also perform many other tasks, for which a pilot's experience is essential. These *extra duties* (often deskwork) are, from the perspective of the Air Force, complements to *normal duties* and therefore vital. But I can demonstrate that direct superiors cannot always supervise the amount of hours a pilot spends on extra duties.[51] These figures are only available to the headquarters, for other statistical evaluations. Therefore the amount of extra duties a pilot performs may be considered as unverifiable and hence to a great extent at his own discretion.[52] The wage effective performance evaluations applied by the Swiss Air Force is

[50] Being myself a pilot of the Swiss Air Force helped me to understand the complex dataset.

[51] According to my analysis, performing more extra duties does not result in a better subjective performance evaluation.

[52] Furthermore pilots may be less intrinsically motivated to fulfil these tasks rather than to perform normal flight duties.

thus primarily based on subjectivity. But this subjectivity may give rise to reciprocal reactions in the field of unobserved efforts: If pilots were to react to changes in their subjective performance evaluations by changes in the accomplishment of extra duties, reciprocity would then prevail.

My results show reciprocal behaviour in the accomplishment of extra duties after a change in the individual subjective performance evaluation. This behaviour however vanishes after one period if the same subjective performance evaluation is issued again. A possible explanation is a strong dependence of effort on expectations. Hence, as long as expectations are met, the provision of unobserved effort seems to be sustained but not enhanced due to reciprocity.

The rest of this chapter is structured as follows: In section two I provide the fundamentals by giving a quick review of the literature. I furthermore provide more details about Swiss Air Force pilots and introduce the applicable theory of Benjamin (2006). Section three renders the propositions. The empirical analysis and the results are discussed in section four. Finally section five is the summary and conclusion of this chapter.

3.2. Fundamentals

3.2.1 Prior Literature

In recent years many theoretical papers on the subject of reciprocal behaviour and fairness try to answer several basic questions (e.g. Akerlof, 1982; Akerlof and Yellen, 1988 and 1990; Fehr and Schmidt, 1999; Falk and Fischbacher, 2000; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Camerer and Fehr, 2006). In addition, a growing literature researches the matter by using experiments to predict human reactions.[53] Field data about human behaviour also confirm reciprocity on a general level. E.g. Krueger and Mas (2004) found in the tire industry that bad working conditions had been retaliated with bad quality production.[54] Field

[53] Some studies in order of their appearance are Fehr, Kirchsteiger and Riedl (1993), Berg, Dickhaut and McCabe (1995), Fehr, Gachter and Kirchsteiger (1997), Fehr and Falk (1999), Fehr and Gachter (2000b), Falk and Gachter (2002), Andreoni, Harbaugh and Vesterlund (2003), Fehr and Schmidt (2004), Fehr and List (2004), Charness (2004) and Gachter, Kessler and Koenigstein (2006).

[54] Other Studies are Katz, Kochan and Gobeille (1983), who found effects off grievance, discipline rates and union-management climate on economic performance, Cappelli and Chauvin (1991) found that a wage premium reduces shirking, Kleiner, Leonard and Pilarski (2002) found effects of strikes and slowdowns on performance in the aircraft manufacturing industry, Bartel et al (2004) correlate a opinion survey to performance in a major bank, Mas and Berkeley (2004) found that labour unrest causes bad quality in construction equipment and Mas

data about individual wages, in contrast, have mostly been evaluated in the aspect of the direct interdependence of incentives and relevant measured performance (e.g. Lazear, 2000).[55] Another avenue of research has been with tournaments, which reflect the purely economic considerations of employees, who evaluate marginal costs versus marginal returns.[56] Even the idea of *efficiency wages* is a mere evaluation of the returns (such as *shirking)* versus the costs (such as loosing a job) (Shapiro and Stiglitz, 1984).[57] Only the study of Gneezy and List (2006) is closely related to this chapter, as they research reciprocal reactions of workers using experimental field data.[58] So there is to my knowledge no study so far which explicitly tries to answer whether subjective performance evaluations affect unobserved effort in long-term employments by means of reciprocity.

3.2.2 Institutional Background

To be able to understand the effects of reciprocity, I will have to step back and focus on the specific work environment of Swiss Air Force pilots. Let me thus discuss their freedom of action, their different duties and their incentive system.

The contracts of Swiss Air Force pilots are kept largely vague when it comes to working hours. It states, "Work according to needs." (own translation of Federal Department of Defence, 2003a:8, Art.19). The government, who wants pilots to protect the airspace above Switzerland, transport passengers and cargo in times of war and peace, defines this 'need'. Transport and protection requests are, for example, in the case of the annual World Economic Forum in Davos, very precise. Pilots are then tasked by their Centre of Operations. A pilot,

---

(2006) researched the effects of final offer arbitrations and found very significant results, which indicate that workers shirk more, and work less passionate if they loos arbitration.

[55] Other studies are e.g. Fox, Scott and Donohue (1993), Rajagopalan (1997) and Fehr and Goette (2005). Even Kahn and Sherer (1990), who researched the effects of a subjective performance assessment on later performance, used the same performance evaluation to measure later effort levels as was used to determine bonuses and merit pay rises.

[56] Some studies are e.g. Coupé, Smeets and Warzynski (2003) and Lavy (2004). Bandiera, Barankay and Rasul (2005) even compare piece-rates to tournaments. There are also studies on the side effects of incentive schemes as e.g. Camerer and Hogarth (1999), Dohmen and Falk (2006), Eriksson and Villeval (2004) or Eriksson, Teyssier and Villeval (2006). Furthermore Krakel (2004) studied the effects of emotions on the performance of incentive systems.

[57] For an early overview of related theories see e.g. Katz (1986).

[58] Also related is the study of Fairis and Alston (1994) who use individual data but as a proxy for the effort level the worker's self-assessment whether his job requires him to work hard. For further overviews see e.g. Prendergast (1999).

however, may tell the dispatcher in advance that he has other duties to perform. He would then be relieved from flying missions if at all possible. At other times, when only few tasks have to be performed, pilots should train themselves to maintain a certain level of proficiency. Again, a pilot might only perform the compulsory minimum training if he has other things to do. This provides pilots a lot of flexibility on how to spend their time. They may either fly a lot or perform other duties mostly at their own discretion. But why does the system allow professional pilots to sit on the ground and what other duties are there to be performed?

Besides various *normal duties* like flying missions, learning or teaching to fly and everyday administrative work, there are many *extra duties*, which nevertheless have to be performed by pilots. A non-complete list of these *extra duties* would include representative work, the organisation of operations, work for superiors or other organisations as representatives of the Air Force, the publication of new manuals, the evaluation and re-evaluation of equipment, work in the field of human resources and IT, pre-screening or flight safety activities, missions abroad and other special projects.[59] All these duties have common factors that only a pilot's special know-how makes his participation crucial and thus complement normal duties. If for example nobody manages the tiny airspace above Switzerland, efficient training is jeopardised. Some of these extra duties are performed by pilots in management positions. To maintain their pilot skills they also fly part-time. Many other extra duties are achieved on the basis of voluntary commitment. But effort-costs are presumably higher for extra duties than for normal duties as many pilots would rather sit in an aircraft than perform deskwork duties.[60] This intrinsic motivation causes an immanent lack of pilots who voluntarily perform extra duties.[61] In such situations of shortages, the performance of extra duties bears a higher relative payoff for the Swiss Air Force than that of normal duties. As a result almost every superior is willing to allow for less normal duty time as, for example, flying missions, in order to let a pilot perform more extra duties. They even permit certain pilots to become less operational and, for example, only fly small aircraft to allow them to do all their extra duties. This trade-off between the performance of normal and extra duties may thus be seen as a typical motivational problem in a multiple-task environment (Prendergast, 1999:22-29), which may be solved by an adequate incentive system. Let me thus turn to this issue.

[59] For a detailed list see Table 21 in the appendix.

[60] This view is further supported by the fact that pilots have been especially pre-screened for their motivation to fly (Noser, 2003).

[61] A related study by Cavalluzzo (1991) researches the value of a command over deskwork for US Air Force pilots.

The salary of a Swiss Air Force pilot consists of two parts: The less interesting part is a fixed wage component that a pilot receives in the light of compensating wage differentials (Borjas, 2002:201-225) as long as he bears the higher risk of death caused by military aviation (Federal Department of Defence, 2003b:1272, Art.2).[62] The more interesting part of his salary depends on his actual hierarchical position and more importantly on a subjective performance evaluation. This performance evaluation, which is of central interest to us, was introduced in November 2000 and became wage-effective two years later. It replaced the previous 3% annual wage increase which pilots received until their wage reached the maximum for their wage bracket, which is defined by their hierarchical position.[63] Now the yearly wage-increment is only affected by the subjective performance evaluation called LOBE and may vary between 0% and 6% of the maximum of one's wage bracket. LOBE, which stands for "Lohnrelevante Beurteilung" and translates as "wage-relevant evaluation" (or literally "to praise"), is a purely subjective performance assessment as suggested by Baker (2000:420).[64] Every year in November the pilot has to agree to individual evaluation criteria his superior has prepared. These criteria may consist of a pilot "showing a strong willingness to fulfil his tasks" or of a pilot "using wisely his freedom of action". The superior then evaluates to what extent the pilot fulfils these criteria. The following October, the pilot then receives the result of his subjective performance evaluation. The result is rated and ranges from 1-5. With an evaluation of 1 the employee would not get any wage increase and may fear lay-off. With an evaluation of 5, the employee would get a wage increase of 6% of the maximum salary of his wage bracket or up to a 12% bonus-payment if he already reached this maximum. It is normally reached after about 20 years of tenure for a normal pilot.[65] If a pilot gets promoted to squadron commander or into another management position, he will each time change into another wage bracket with an approximately 5% higher maximum. He once gets a 5% salary increase and then falls back into the LOBE-cycle but gets his salary rises in reference to the maximum of his new wage bracket. Even though it would seem attractive to, at first, change wage brackets, the fact that few takers for managerial positions have been available in recent

[62] This risk is not to be confused with the economic risk-sharing problem between employers and employees (see e.g. Bloom and Milkovich, 1998).

[63] These wage increments had been a pure reflection of tenure as described in e.g. Prendergast (1999:45-51) as deferred compensation.

[64] For a discussion of the advantages and disadvantages of such systems in different governments see e.g. Eisenberg and Ingraham (1993).

[65] If good evaluations and therefore high wage increases appear at an early stage, the system can be thought of as one with long-term earning effects as the merit system in Kahn and Sherer (1990:112S).

years (Air Transport, 2007:3), a respective tournament may be excluded. A plausible reason may be the increased amount of extra duty that may bear higher effort-costs than the 5% difference in salary may be able to compensate.

To prevent the sum of all salaries from rising due to a "leniency bias" (Prendergast, 1999:30), a forced distribution is established (Swiss Government, 2001:21, Art.49; Head of the Armed Forces, 2004). All the same, the LOBE-system can not be considered as a tournament, because everybody is working on his individual criteria and has no idea of the results of his subjective performance evaluation relative to others. But this result may serve as a perfect proxy for real salary changes as it is a standard evaluation independent of inflation and wage brackets that may bear income effects (Prendergast, 1999:50).

With all of this in mind, LOBE might seem to easily resolve the motivational problems within the multi-tasking environment described earlier: Even though pilots may have lower effort-costs for normal duties, they should be motivated to fulfil both duties to receive an adequate subjective performance evaluation. And pilots are indeed required to declare the allocation of their working hours to different tasks on a daily basis. Nevertheless, this data is only available to the headquarters that uses them exclusively to statistically evaluate the different tasks. This means that the amount of duties performed is not only unverifiable but also unobserved by the direct superior. Furthermore some superiors sometimes hardly ever see their subordinates as the pilots are deployed to different airbases on an irregular basis.[66] Consequently subjectivity may play a major role in this evaluation system.[67] How then are pilots motivated to perform extra duties since the costs for their accomplishment seems to be higher than for those of normal duties? One possible means may be the effects of reciprocity: An employee might reward the result of a subjective performance evaluation if the superior appreciates the employee's work.

Since, unlike direct superiors, I was given access to the operational data from the former headquarters, allowing me to pursue the research of this idea. But allow me to start by reviewing a theory that may be able to predict pilot behaviour.

---

[66] I hence deviate from the commonly assumption that some effort is observable but not verifiable (Gibbons, 1998:121). This fact would rule out the point of Brown, Falk and Fehr (2002) who showed that repeated interactions might overcome the leaks of incomplete contracts. If however the leak is not observed by the superior, their findings may not be applied to the described situation.

[67] This might be especially true because superiors are never fully objective when issuing a subjective performance evaluation (Schettgen, 1996:261, Fallgatter, 1999:84-87).

3.2.3 Theory

Out of the many theoretical papers on the subject of reciprocal behaviour and fairness, I decided to use a theoretical work by Benjamin (2006), who applied the well-known theory of inequity aversion by Fehr and Schmidt (1999) to the labour market situation. Here is a quick review of the theory and their later development:

Fehr and Schmidt (1999:822) define the utility of person *i* in a two-player game of *i* and *j* as follows:

$$U_i(x)=x_i - \alpha_i max\{x_j-x_i,0\} - \beta_i max\{x_i-x_j,0\},\ i \neq j \quad (1)$$

In equation (1), x_i and x_j are the shares the two players get. The second term reflects person *i*'s utility loss from disadvantageous inequality, while the third term measures the loss from advantageous inequality (Fehr and Schmidt, 1999:822). Therefore α_i might reflect person *i*'s motivation to show negative reciprocity and β_i to show positive reciprocity.[68]

To use this theory in my environment I benefit from its later development by Benjamin (2006), who applied it to the general situation of firms and workers. He transformed α and β in the following way (Benjamin, 2006:32):

$$\phi = \alpha + \beta\ ,\ \gamma = \frac{\beta}{\alpha + \beta} \quad (2), (3)$$

Now ϕ reflects the weight an employee puts on fairness, while γ reflects the importance of fairness for the worker when he is better off and $(1-\gamma)$ when he is worse off.
In his theory, the firm's payoff π^F and the worker's payoff π^W depend on the offered wage *w* and the later chosen effort level *e* (Benjamin, 2006:6):

$$\pi^F(w,e)=y(e)-w, \text{ while } y'>0,\ y''<0 \quad (4)$$

$$\pi^W(w,e)=v(w)-c(e), \text{ while } v'>0,\ v''<0 \text{ and } c'>0,\ c''>0 \quad (5)$$

In (4) and (5) $y(e)$ is the firm's production function, $c(e)$ the worker's cost function and $v(w)$ the worker's function for the benefit of higher compensation.
Furthermore Benjamin (2006:7-8) defines surplus payoffs of a firm and a worker as:

$$\tilde{\pi}^F \equiv \pi^F(w,e) - \hat{\pi}^F \quad (6)$$

$$\tilde{\pi}^W \equiv \pi^W(w,e) - \hat{\pi}^W \quad (7)$$

[68] Fehr and Schmidt (1999:853) claimed in their original paper that their model is also able to predict reciprocal behaviour. Hence instead of measures of inequality aversion, the parameters may also be regarded as measures of reciprocity.

In (6) and (7), $\hat{\pi}$ are the reference payoffs of the firm and worker. These will play a major role in the further analysis.

Benjamin (2006:8-9) now assumes that the worker maximises his utility:

$$U^W = \pi^W + \phi[-\gamma max\{\tilde{\pi}^W - \tilde{\pi}^F, 0\} - (1-\gamma)max\{\tilde{\pi}^F - \tilde{\pi}^W, 0\}] \quad (8)$$

Analogical to (1) the long second term in equation (8) reflects the workers care for fairness. To assess which effort level the worker chooses, Benjamin (2006:15) first postulates a reference effort level for the case that the firm's and the worker's surplus payoffs are equal:

$$e^{fair}(w) :\equiv \arg\max_e f\,[\pi^W(w,e) - \hat{\pi}^W,\ \pi^F(w,e) - \hat{\pi}^F] \quad (9)$$

This means that a fair effort level satisfies (10):

$$\pi^W(w,e^{fair}) - \hat{\pi}^W = \pi^F(w,e^{fair}) - \hat{\pi}^F \quad (10)$$

This fair effort level is strictly increasing in wage *w* because the worker's surplus payoff is rising while the firm's surplus payoff is decreasing in *w*. The only way to equalize surplus payoffs for the worker is to choose *e* to match e^{fair}.[69] The question now is whether the worker wants to choose e^{fair} at all:

Rearranging (8) depending on the postulated reference effort level e^{fair}, the worker's utility is (Benjamin, 2006:36):

$$U^W = \begin{cases} \hat{\pi}^W + (1+\phi(1-\gamma))\tilde{\pi}^W(w,e) - \phi(1-\gamma)\tilde{\pi}^F(w,e) & \text{if } e > e^{fair}(w) \\ \hat{\pi}^W + (1-\phi\gamma)\tilde{\pi}^W(w,e) + \phi\gamma\tilde{\pi}^F(w,e) & \text{if } e \le e^{fair}(w) \end{cases} \quad (11)$$

The first term in (11) is easy to analyse: U^W is always decreasing in e, since $\partial\tilde{\pi}^W(w,e)/\partial e < 0$ and $\partial\tilde{\pi}^F(w,e)/\partial e > 0$. Hence a utility maximising worker will always reduce his effort level until it reaches e^{fair}. The second term in (11) is strictly concave in *e* (Benjamin, 2006:36). It is however only increasing in *e* until e^{fair} if the following condition (12) is met (Benjamin, 2006:37).[70]

$$-(1-\phi\gamma)c'(e^{fair}) + \phi\gamma y'(e^{fair}) \ge 0 \quad (12)$$

Rearranged, (12) solves for ϕ (Benjamin, 2006:32):

$$\phi \ge \frac{1}{\gamma\left(\frac{y'(e^{fair})}{c'(e^{fair})} + 1\right)} \quad (13)$$

[69] For further details see Benjamin (2006:15).

[70] This follows because $\frac{\partial U^W(w,e)}{\partial e} = (1-\phi\gamma)\frac{\partial\tilde{\pi}^W(w,e)}{\partial e} + \phi\gamma\frac{\partial\tilde{\pi}^F(w,e)}{\partial e}$ and $\frac{\partial\tilde{\pi}^W(w,e)}{\partial e} = -c'(e)$ and $\frac{\partial\tilde{\pi}^F(w,e)}{\partial e} = y'(e)$.

This means that the utility-maximising worker only increases his effort level until it reaches the postulated e^{fair} if his personal ϕ and γ are strong enough. This requires the worker to care for fairness for the case that he is better off than the firm. The stipulation of (13) is further rather true if the firm's increase in payoff π^F from an increase in e^W is higher than the worker's increase in cost c^W form the same increase in effort level e^W. Hence on the condition of (13), the worker's chosen effort level directly depends on his wage and his reference transaction.[71] To assess whether this is actually the case for the Swiss Air Force pilots, I will merge the collected information and theoretical knowledge and proceed by the formulation of adequate propositions in the next section.

3.3. Propositions

To be able to propose a pilot's decision-process, I start by evaluating his payoff function without considerations of reciprocity. It seems reasonable to assume that also a pilot's payoff function follows (5), as he might like to earn more while bearing smaller costs for the effort involved. While wages seem to directly influence the pilot's benefits, it is more difficult to analyse his effort-costs. Apart from the difference in effort-costs for different types of duties, there seems to be two more aspects to effort: On the one hand, costs may arise from the way a pilot does his duties and on the other hand, they may arise from the amount of time he spends for these duties. As the subjective performance evaluation rewards the way the duties are performed, it may be assumed that a pilot willingly bears some costs for this kind of effort, because this behaviour may maximise his payoff.

But what about the amount of time a pilot spends on different duties? If no reward is granted for this kind of effort because it is actually unobserved, standard incentive theory predicts a pilot will bear minimum effort-costs in this sector and hence not work at all.[72] But this corner-solution is unlikely. I believe that even though pilots may have self-control to a great extent, a pilot that is never around tends to attract attention. This means that one should not assume that duty-time is totally unobserved by superiors. Hence the liberty of a pilot to choose the amount of total working time is only true within a certain range. What options are left to help

[71] For further details and the firm's optimal choice of wage see Benjamin (2006).

[72] Due to this gap in LOBE, the performance enhancing effects of such an incentive system might be lost. Performance enhancing effects have been shown by e.g. Lazear (2000:1346) for piece rate workers or by e.g. Kahn and Sherer (1990:118Sf) for the case of managers in terms of the influence of bonuses on their later performance.

minimize effort-costs that grant no rewards? The pilot might optimise the different shares of duties he performs: Effort-costs for normal duties might be reasonably low, since a pilot likes to fly. Effort-costs for time spent on extra duties however appear to be higher than for normal duties. Therefore even a substitution of the two kinds of duties implies lower total effort-costs for a pilot. According to these considerations, standard incentive theory predicts that a pilot chooses a minimum amount of extra duties as soon as he perceives that the additional performance of extra duties remains unrewarded.[73] Having however seen that the performance of extra duties is very important to the firm, considerations for fairness and reciprocity might become relevant. Hence a different reasoning may apply. To evaluate this issue, I will now make use of the introduced theory by Benjamin (2006).

It is important to bear in mind that in Benjamin's (2006) theory effort comes after wage unlike in an incentive theory. But the time spent on extra duties does not influence the subjective performance evaluation. Hence wage does not come after this kind of effort. The evaluations seem to be rather based on subjectivity, which implies that the pilot probably does not fully deserve the wage he earns. If a pilot believes his evaluation depends, at least to some extent, on unverifiable reasoning, he might show reciprocity. Hence, effort might come after wage as in Benjamin's (2006) theory.

To apply considerations of reciprocity, I also have to consider the payoff of the Air Force. The payoff of such a governmental organisation is however very difficult to measure. But I may assess its production function. The Air Force produces "safety and transportation" as a primary product. This of course is mainly done by the performance of normal duties. How then do extra duties affect this product? Because the accomplishment of extra duties is complementary to normal duties, an additional performance of extra duties should improve the productivity of the system. This is the case even if normal duties are substituted as long as lack of pilots wanting to perform extra duties prevails. The increased productivity results in a better accomplishment of the missions of the Air Force and therewith potentially increases its payoff. As I am only going to consider relative changes it suffices to learn that this firm's payoff $\pi^W(w,e)$ is indeed according to (4) increasing in e (the amount of extra duties performed) and decreasing in w (the wage of a pilot) because this reduces the available money of the Air Force's limited budget.

A quick conclusion would now be that a pilot chooses his unobserved effort level, i.e. the amount of extra duties he performs, according to his salary. This might be true for the first

[73] This situation may be seen as equivalent to the "multi-tasking problem" described by Prendergast (1999:23).

two periods of his employment.[74] Thereafter further considerations might apply: As discussed earlier, part of the subjective performance evaluations, which primarily affect the later wage changes of a pilot, might actually not be earned by the way he performed his duties before. Based on the subjectivity of the evaluation I therefore make the following important assumption: A pilot tries at least to do his job as good as the previous year. Hence he expects at least the same evaluation and therefore the same wage increment as before.[75] If he chooses in the first two periods of his employment a "fair" unobserved effort level, he might thereafter always compare the received evaluation to that of the year before.[76] Thus, if his expectations are met, he exerts the same "fair" unobserved effort level than before.[77] This assumption might be challenged because a pilot might have tried to perform his duties better than before and therefore expect a better evaluation than last year. I, however, believe that even for this case he is positively surprised if he actually receives a better evaluation because the superior not only appreciated his endeavours although he may not have seen him much, but also

[74] The initial "fair" amount of extra duties depends on the reference payoffs the pilot uses for himself and the Air Force. They may vary from pilot to pilot as not all pilots have the same education and outside options. Let me however assume that he initially chooses to perform this "fair" amount of yearly extra duties (at least in his view, which is relevant for this case).

[75] This assumption is motivated by Kahneman, Knetsch and Thaler (1986:730): "The current wage of an employee serves as reference for evaluating the fairness of future adjustments of that employee's wage." Also the literature about "psychological contracts" supports the importance of expectations (see e.g. Herriot, Manning and Kidd 1997).

[76] Albeit he already receives at the end of the first period a subjective performance evaluation, he may not be able to judge its fairness as he still lacks a point of reference.

[77] This assumption also takes care of the problem of deferred compensations (Prendergast, 1999:45-49), which means that wages in many firms are normally low at the beginning and high at the end of a career. If the fair effort level would simply depend on absolute wage like in (10), it would have to follow the rising wage of such a compensation system. But deferred compensations are mainly meant to bind a worker to the firm (Prendergast, 1999: 45). Thus the worker expects these promised compensations and does not change his initially chosen yearly effort level as long as the promises are kept. For the case of the Swiss Air Force this would mean that as long as "normal" wage increments are granted, the worker performs the same yearly amount of extra duties. Further explanations for this phenomenon are given by e.g. Backes-Gellner, Lazear and Wolff (2001:1-72). The rising wage of employees may also reflect their internal knowledge and their consequently boosted productivity. For the case of the Swiss Air Force I simplify this picture by assuming that a worker, although more productive according to his experience level, still bears the same costs to his effort level.

choose him for a better evaluation within the forced distribution of all evaluations.[78] If the pilot however receives a worse evaluation than before, he may be disappointed.

To implement this dynamic situation into the model, I make use of the reference payoffs: I define the reference wage $\hat{w}$ in the worker's reference payoff $\hat{\pi}^W \equiv \pi(\hat{w}, \hat{e})$ as the expected wage. As there is almost no outside option for the firm for native military aviators, I assume that the firm's reference transaction depends as well on the expected wage.[79] I define this expected wage $\hat{w}$ as the pilot's last wage plus his expected wage increment, which I take to be the same as his last wage increment.[80]

To be able to assess a fair unobserved effort level in terms of the amount of extra duties performed, I also have to define a respective reference effort level $\hat{e}$. Again, neither the firm nor the worker has an easily achievable outside option that could be used as a reference transaction.[81] Hence I rather use the worker's labour market experience in terms of unobserved effort levels inside the firm. For this experience I see two possible alternatives: Either I assume that the reference unobserved effort level $\hat{e}$ is always the pilot's *last year's unobserved effort level* e_{t-1} that reflects his most recent labour market experience inside the firm, or his personal *standard unobserved effort level* e_s that potentially reflects his initial "fair" unobserved effort level inside the firm or his long-term labour market experience.[82] The first version implies that pilots always provide unobserved effort levels in reference to the previous year; the second version implies that pilots do not especially consider the previous year's unobserved effort level, instead focus on their long-term experience. As we will later see, the first version simply proposes that a pilot adapts his unobserved effort level constantly to his performance evaluation while the second version entails that a pilot has to be newly motivated by surprises every period. Otherwise he settles back to his standard unobserved effort level even if he receives constantly very good performance evaluations. For both cases I will however only make predictions in relations to previous effort levels.

---

[78] If the superior does not render a better evaluation for this case, the employee might of course be disappointed. I am however not able to assess these cases but believe, after having talked to some pilots, that my assumption bears some realistic features.

[79] This seems at least probable for the pilot's point of view that is relevant for this case.

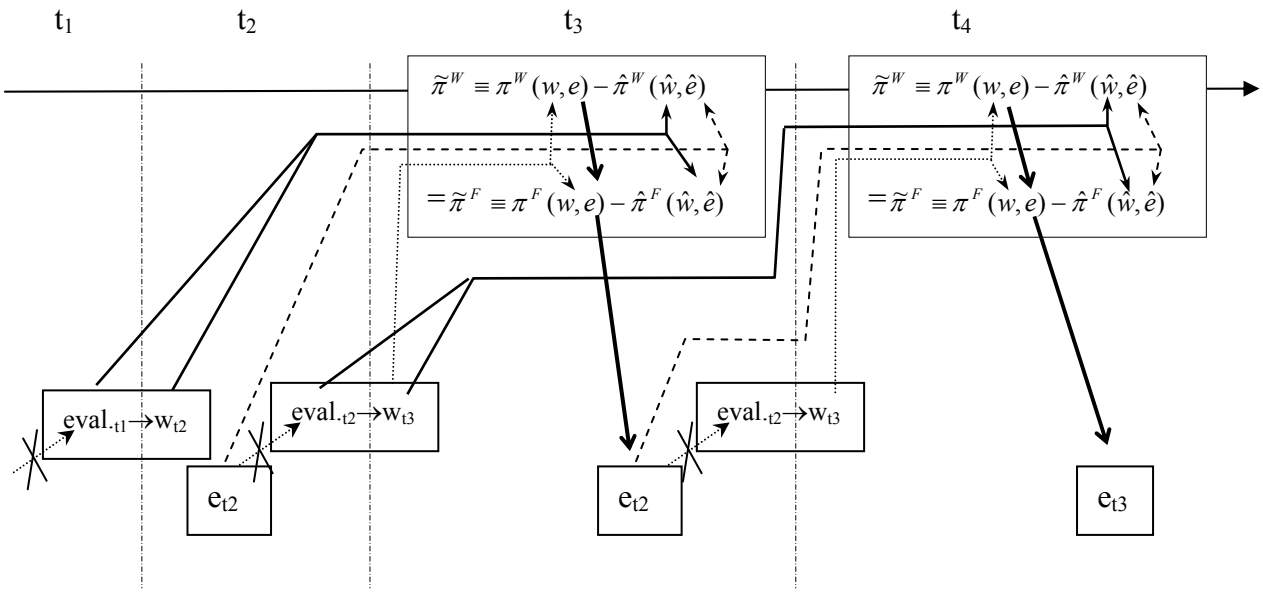
[80] This view is further supported by the study of Mas (2006). The police officers final offer obviously served them as a reference point for expected wage.

[81] For the availability of outside options for Swiss Air Force pilots see Hedinger (2004).

[82] These arguments are in line with Benjamin (2006:3) who reasons that the reference transaction may depend on current and past market rates and the worker's recent labour market experience.

But let me give a more detailed analysis to be able to make propositions for the available data of 4 periods of LOBE-Data and 2 periods of extra duty time, to which I will refer to from now as "unobserved effort level" *e*: I start by analysing the simpler case where I assume that the worker's reference effort level is his *last year's unobserved effort level* e_{t-1}. I show the case for my available data in Figure 9 where the x-axis is the timeline with periods t_1, t_2, t_3 and t_4.

Figure 9: The four period application of Benjamin's (2006) model to the situation of the Swiss Air Force with expected wage and reference effort levels *as last years unobserved effort level*.



Notes: own drawings

As visualized by a crossed out arrow between each unobserved effort level *e* and the evaluation at the end of a period, it is important to keep in mind that the unobserved effort level (the amount of extra duties) does not influence the evaluation at the end of a year and therewith the wage in the next year. Let me explain the figure by an example. In period t_1 the pilot chooses the unobserved effort level according to his appraisal of a fair unobserved effort level in reference to wage and working conditions. The superior evaluates the work of this pilot in t_1, for instance, as average (independent of last year's unobserved effort level). This directly affects w_{t2}. If this is the first evaluation the pilot receives in his carrier, he most probably accepts it and chooses e_{t2} again according to his initial appraisal of a fair unobserved effort level. But the evaluation at the end of t_1 gives rise to an average expected wage $\hat{w}_{t3}$ in period t_3 because the pilot expects the same evaluation at the end of t_2 again. The respective

coherence is displayed by the solid arrow. As discussed before, I assume for now that the reference effort level $\hat{e}_{t3}$ is equivalent to the unobserved effort level e_{t2} as shown by the dashed arrow. If in period t_2 the superior appreciates the employee's work more than expected, the good evaluation of t_2 gives rise to a high wage in period t_3 ($w_{t3} > \hat{w}_{t3}$). This wage w_{t3} now enters along the doted arrow the fairness assessment function of the pilot. If in t_3, the pilot does not change his unobserved effort level but puts forth again $e=\hat{e}=e_{t-1}=e_{t2}$ (along the bold arrow), the surplus payoff for the worker $\tilde{\pi}^W \equiv \pi^W(w,e) - \hat{\pi}^W(\hat{w},\hat{e})$ is positive because $\pi^W(w,\hat{e}) > \pi^W(\hat{w},\hat{e})$. But the surplus payoff for the firm $\tilde{\pi}^F \equiv \pi^F(w,e) - \hat{\pi}^F(\hat{w},\hat{e})$ is negative because $\pi^F(w,\hat{e}) < \pi^F(\hat{w},\hat{e})$. Hence if $w > \hat{w}$, then $e^{fair} > e_{t-1}$ must be true. The worker is therefore expected to increase his unobserved effort level in t_3 and choose $e^{fair} > e_{t2}$ if his utility of the second term in (11) is increasing in e. This is the case if (13) holds for these particular pilots, which is very plausible because according to chapter two Swiss Air Force pilots show high values for ϕ and γ.[83] If at the end of t_3 the employer again renders the same good qualification (independent of the unobserved effort level in t_3), then $w_{t4} = \hat{w}_{t4}$. Hence $\pi^W(w,\hat{e}) = \pi^W(\hat{w},\hat{e})$ and $\pi^F(w,\hat{e}) = \pi^F(\hat{w},\hat{e})$ if the worker selects again $e=\hat{e}=e_{t-1}$. Therefore if $w = \hat{w}$, then $e^{fair} = e_{t-1}$.

For the negative case, if in period t_3 the *expected* $\hat{w}$ is not granted, surplus payoff $\tilde{\pi}^W$ is negative while $\tilde{\pi}^F$ is positive if the worker chooses the same unobserved effort level e as in t_2. Hence $e^{fair} < e_{t-1}$ is true. As the utility from the first term of (11) is always decreasing in e, the worker decreases his unobserved effort level and chooses $e^{fair} < e_{t2}$. If at the end of period t_3 again the same evaluation is issued, then $w = \hat{w}$, and therefore the worker will choose $e^{fair} = e_{t-1}$. As we can see, the evaluation affects the unobserved effort level in the long run. Every chosen unobserved effort level is also the starting point for the path of all of a pilot's later unobserved

[83] At least 71% of the pilots in chapter two show values for β that are higher than 0.3. As (13) transforms for β into $\beta \geq \left(\frac{y'(e^{fair})}{c'(e^{fair})} + 1 \right)^{-1}$ it all depends on the relationship of y' and c'. Even though y' must be high as a permanent lack of pilots wanting to perform extra duties persist, I can not make any assumption of whether it is higher than the marginal effort-cost for the pilot. High marks for ϕ and γ seem also supported by two further studies: Henrich et al (2001:76) found in their experiments that people who have a high payoff to cooperation in their everyday life are more cooperative than others. This might very well apply to pilots. Hennig-Schmidt, Rockenbach and Sadrieh (2005:11f) finally found in a laboratory experiment that workers do show reciprocity if they are informed about the employer's surplus, which is probably the case for extra duties of pilots.

effort levels.[84] Hence unobserved the effort levels e_{t1} and e_{t2} are either the unobserved effort levels he chose in the first and second period of his employment or have been predetermined by the same scheme as e_{t3} and e_{t4} by previous data. These findings are the grounds for my first proposition. I will however try to further simplify the coherences to make the proposition easier to test in the real life environment of my data.

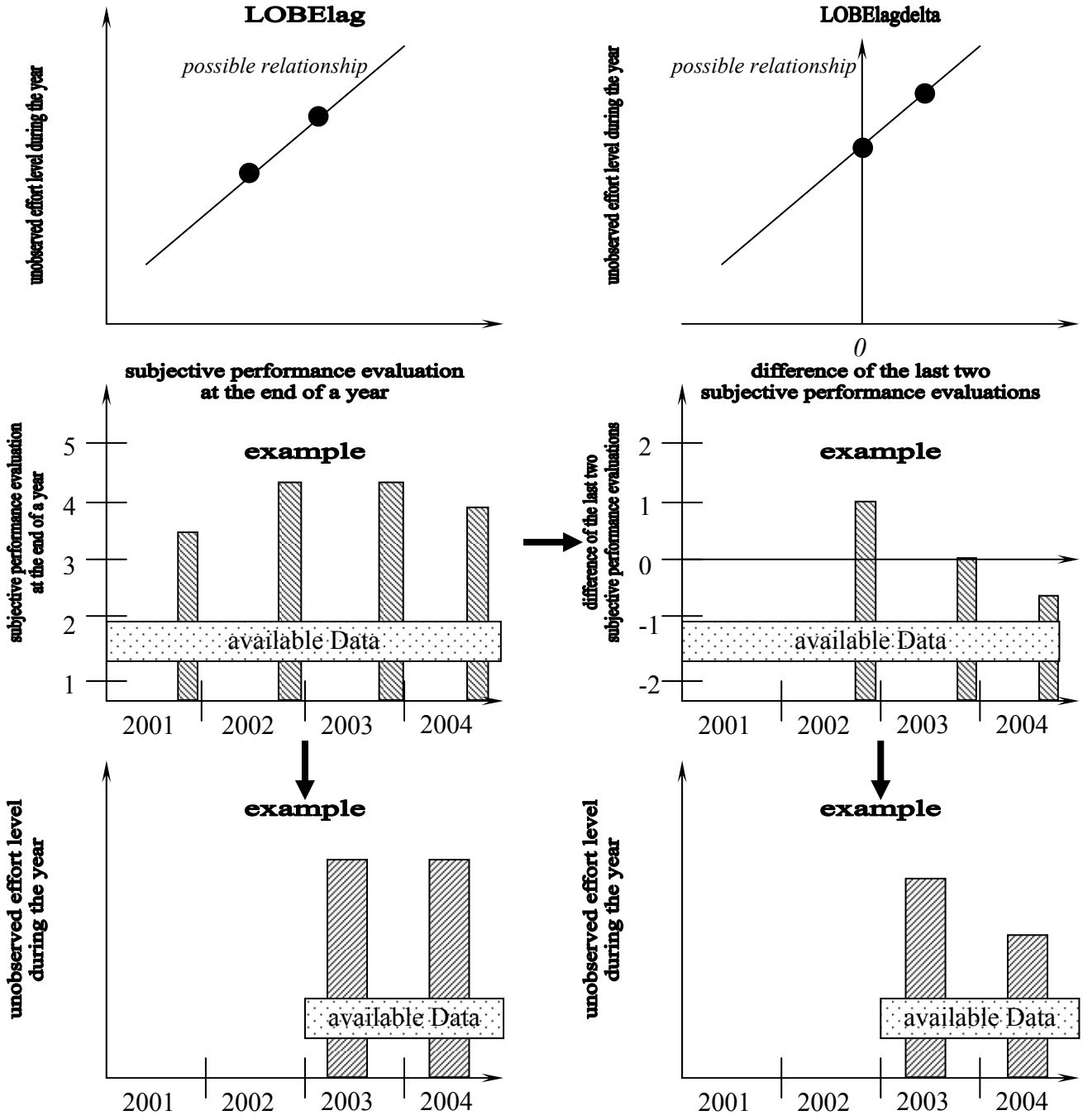
I recapitulate that the pilot's expected wage in t_3 is defined by his wage in t_2 plus the wage increment according to his subjective performance evaluation in t_1, i.e. $\hat{w}_{t3} = w_{t2} + f(eval_{t1})$. The pilots wage in t_3 depends on his subjective performance evaluation in t_2 and on his wage in t_2, i.e. $w_{t3} = w_{t2} + f(eval_{t2})$. Because the surplus payoff of the worker $\tilde{\pi}_{t3}^{W} \equiv \pi_{t3}^{W}(w,e) - \hat{\pi}_{t3}^{W}(\hat{w},\hat{e})$ positively depends on the difference between w_{t3} and $\hat{w}_{t3}$ it must also positively depend on the difference between the last two evaluations ($eval_{t2}$-$eval_{t1}$). This means that a change in the evaluation from t_1 to t_2 implies a change in the unobserved effort level from t_2 to t_3.

In the left column of Figure 10 I try to visualize these circumstances in simple graphs. I show in the top left panel on the x-axis the evaluation of t_{-1}. On the y-axis I show the chosen unobserved effort level in t. For the previous example this would mean that the pilot in t_1 received a medium evaluation and in t_2 choose to apply a medium unobserved effort level that draws a first point in, for example, the middle of the graph. At the end of t_2 he received a better evaluation, hence his expectations are surpassed and he chooses to apply in t_3 a higher unobserved effort level than in t_2. This draws a second point in the graph which is higher and more to the right than the first point. In period t_4 nothing would happen, as the pilot gets the same evaluation and hence chooses to apply the same unobserved effort level as before. If I connect these two points, a positive dependence between the last received result of the subjective performance evaluation and the chosen unobserved effort level materializes. Even though I do not know the exact coherences, I visualize this by a simple linear dependence in the top left panel of Figure 10. I am slightly more careful when I formulate the following simple proposition:

- ***Proposition 1:*** *The time a worker spends on unobserved tasks is a function of his most recent result of his subjective performance evaluation.*

[84] I do not believe that a pilot strategically chooses his effort level, as it will enter his decision function in the next period. I rather believe that he "rewards" fair wages with fair effort levels from period to period.

Figure 10: Interdependence of subjective performance evaluations, time and unobserved effort levels.



Notes: own drawings

In the second panel of the left column of Figure 10 a critical example for available data is displayed: I show a possible course of the results of an individual evaluation. It starts with 3.5 for the period of 2001, continues with twice an evaluation of 4.5 for 2002 and 2003 and ends with an evaluation of 4.0 for the period of 2004. I accordingly predict for the years 2003 and 2004 the highest unobserved effort level, which I show in the left panel of the third row of Figure 10.

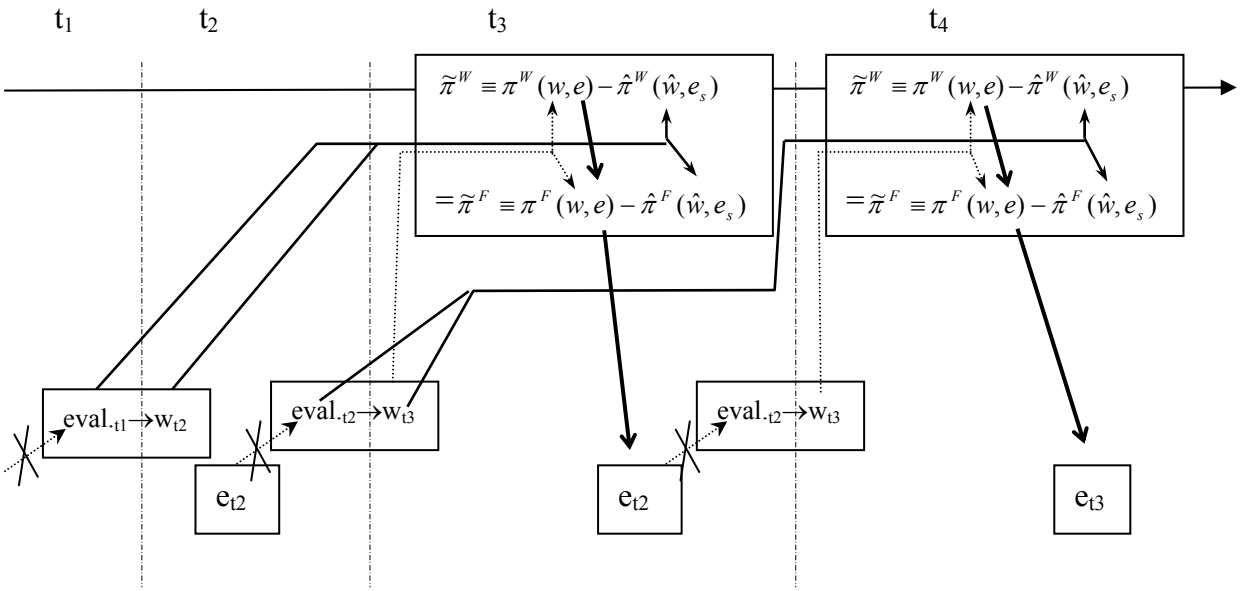
Before turning to the next proposition I give a quick preview of a possible consequence of proposition one: If it is confirmed, the motivation problem in the multiple task environment is solved. Unobserved tasks are not endangered as they are accomplished by the workers who reciprocate the receipt of good subjective performance evaluations. Even though it is up to the worker to exert an initially perceived fair unobserved effort level, this implies that in the case that unobserved effort levels are too low, the average evaluation may be raised and consequently the problem solved.

I now turn to the analysis of the slightly more complex situation where I use a *standard unobserved effort level* e_s which stands, for example, for the initial "fair" unobserved effort level of an individual pilot. I am going to use it as the reference effort level $\hat{e}$.[85] I use Figure 11 to visualize the situation. The x-axis is again the timeline with periods t_1, t_2, t_3 and t_4. The two only differences to the already described situation in Figure 9 are the missing dashed arrows connecting the unobserved effort level in one period with the reference effort level in the next period and the substitution of $\hat{e}$ by e_s in the payoff-formulae. Let me explain the new situation by the same examples I used before. If I assume that the pilot chooses in the second period $e_{t2}=e_s$, then no difference appears to the first situation where $eval_{t2} > eval_{t1}$, as the pilot simply chooses in the third period $e_{t3}=e^{fair} > e_s= e_{t2}$. The difference only shows in the fourth period: If at the end of period t_3 he receives the same evaluation as at the end of t_2, then $w_{t4}= \hat{w}_{t4}$. If in t_4, the pilot does not change his unobserved effort level but puts forth again $e = e_{t3} > e_s$, the surplus payoff for the worker $\tilde{\pi}_{t4}^{\,W} \equiv \pi_{t4}^{\,W}(w,e) - \hat{\pi}_{t4}^{\,W}(\hat{w},\hat{e})$ is negative because $\pi_{t4}^{W}(\hat{w},e) < \pi_{t4}^{W}(\hat{w},e_s)$. And the surplus payoff for the firm $\tilde{\pi}_{t4}^{\,F} \equiv \pi_{t4}^{\,F}(w,e) - \hat{\pi}_{t4}^{\,F}(\hat{w},\hat{e})$ is positive because $\pi_{t4}^{F}(\hat{w},e) > \pi_{t4}^{F}(\hat{w},e_s)$.

[85] It is possible that a pilot changes this initial "fair" unobserved effort level as his individual reference transactions might change. What is important to consider is the difference to the first case, where the reference transaction was simply last year's unobserved effort level, while in this second case the reference effort level is basically independent of last year.

Hence if $w=\hat{w}$, and $e_{t-1}>e_s$ then $e^{fair}<e_{t-1}$ because if $w=\hat{w}$, then $e^{fair}=e_s$ is always true. As the utility from the first term of (11) is always decreasing in e the worker decreases his unobserved effort level in t_4 and chooses $e^{fair}<e_{t3}$. This he does in contrast to the previous case where the reference effort level was e_{t-1} and he therefore chose an unobserved effort level equivalent to e_{t-1}.

Figure 11: The four period application of Benjamin's (2006) model to the situation of the Swiss Air Force with expected wage and reference effort levels *as an individual standard unobserved effort level.*



Notes: own drawings

Let me also explain the negative case: The pilot in the third period had chosen $e_{t3}<e_s=e_{t2}$ because his expectations had been spoiled ($eval_{t2}<eval_{t1}$). At the end of t_3 he receives now the same evaluation as at the end of t_2, which means $w_{t4}=\hat{w}_{t4}$. If in t_4, the pilot does not change his unobserved effort level but puts forth again $e=e_{t3}<e_s$, the surplus payoff for the worker $\tilde{\pi}^W \equiv \pi^W(w,e) - \hat{\pi}^W(\hat{w},\hat{e})$ is positive because $\pi^W(\hat{w},e)>\pi^W(\hat{w},e_s)$. And the surplus payoff for the firm $\tilde{\pi}^F \equiv \pi^F(w,e) - \hat{\pi}^F(\hat{w},\hat{e})$ is negative because $\pi^F(\hat{w},e)<\pi^F(\hat{w},e_s)$. Hence if $w=\hat{w}$ and $e_{t-1}<e_s$ then $e^{fair}> e_{t-1}$ must be true. The worker is therefore expected to increase his unobserved effort level in t_4 and choose $e^{fair}= e_s$ if the utility of his second term in (11) is increasing in e. This is the case if (13) holds for these particular pilots, which is very plausible as discussed earlier. Again, the situation is different to the case that the reference effort level was $\hat{e}=e_{t-1}$. Now the fair unobserved effort level does not so much depend on the path of the

unobserved effort levels in previous periods. If an employee perceives a certain unobserved effort level as fair, he only changes it for a short time if his expectations are un-matched. Hence this time the predetermination of the unobserved effort level in t_2 is less important. These findings are the grounds for my second proposition. Again I will try to further simplify the coherences to make the proposition easier to test in a real life environment.

I recapitulate that the worker only exerts higher unobserved effort levels, if his expected wage is exceeded and only lower unobserved effort levels if his expectations are belied. If not, he does not change or even returns to his standard unobserved effort level e_s. This means that instead of keeping up a high unobserved effort level when he receives consecutive high evaluations he quickly accustoms to the good evaluations and chooses again his standard unobserved effort level.[86] But that also means that he accustoms fast to bad evaluations and already in the second period after the deception chooses again his "standard" unobserved effort level.

Figure 10 on page 47 visualizes these circumstances in the right column. In the top right panel I show on the x-axis the difference of the last two subjective performance evaluations. On the y-axis I show again the unobserved effort level. In my first example a worker in t_1 received a medium and in t_2 a better evaluation. As therefore his expectations were surpassed, he chooses in t_3 a high unobserved effort level that draws a first point, for example, at the top right of the graph. At the end of t_3 he received again the same good evaluation. Hence his expectations are not surpassed anymore and he chooses in t_4 his standard unobserved effort level that is lower than e_{t3}. This would draw in the graph a second point below and to the left of the first point. Connecting these points draws again a direct dependence between the two axes. This time however the unobserved effort level depends on the difference of the last two subjective performance evaluations and not just on the last evaluation. Even though I draw a linear dependence of the two values I am aware of the unknown exact relationship between them. I accordingly formulate the second simple proposition for the actual case:

- ***Proposition 2:*** *The time a worker spends on unobserved tasks is a function of the difference between his most recent subjective performance evaluation and the one before that.*

[86] This assumption is motivated by Gneezy and List (2006:8), who found positive reciprocity to vanish over time in a field experiment. The view is further supported by purely psychological papers as e.g. Gilbert et al. (1998) who research the "durability bias" as a restricted period of time that negative feelings endure. This could very well apply to reciprocity as well. Finally it is an old view expressed by well-known philosophers that we quickly accustom to better living standards (see e.g. Schopenhauer, 1922).

In the second row of Figure 10 I proceed with the same example of real data as explained before: The right panel of row two now shows the differences of the last two results of the subjective performance evaluations of the same pilot as in the left panel of row two. As at the end of 2002 he received a better evaluation than in 2001, the difference is positive. He hence gets a better evaluation as he expected and might show positive reciprocity by a high unobserved effort level in 2003. This is displayed in the right panel of row three. At the end of 2003 however he received an evaluation that was at the same level than before and therefore as expected, hence he is not positively surprised and does not show an especially high unobserved effort level in 2004. He therefore exerts a lower unobserved effort level than in 2003.

With this proposition the question of the multi-tasking problem remains partly unanswered. If it is confirmed, I might be able to predict who is increasing and who is decreasing his level of unobserved effort for one period. It does, however, not seem as if the workers are motivated over a longer period by the subjective performance evaluations to provide more than their usual level of unobserved effort.

Comparing both propositions, it seems possible that both of them may get confirmed: If the first proposition augments the standard unobserved effort level, the pilot in Figure 10 might spare no unobserved effort in 2003 and apply a high effort level in 2004, which is higher than the one in 2002.

As mentioned before, for the Swiss Air Force pilots I consider extra duties as unobserved effort not only because they seem to be unobserved but also because they seem to be the perfect field for reciprocity. Because the performance of normal duties is also unobserved, I now turn to the question whether the pilots perform extra duties additionally to their other duties, or if they substitute extra for normal duty. For the case that they want to perform less extra duties, I already answered the question, as pilots may not be able to perform less than a certain amount of total duty. For the case that they want to perform more extra duties, the pilots have two options: Either they bear the full marginal costs of extra duties if they do them additionally or they just bear the difference between the marginal costs of extra and normal duties if they substitute one for the other. Substitution may therefore only be sub-optimal if the marginal costs for normal duties are negative. Even though it is possible that pilots enjoy flying a lot for the first hour it seems very probable that they enjoy the marginal flying-hour less than the first one. Consequently also for normal duties, the marginal costs must be

positive. Hence pleased pilots are expected to do more extra duties but not more total duties. I therefore propose the following:

- ***Proposition 3:*** *Pilots perform additional unobserved tasks by substituting extra duties for normal duties and not by performing the extra duties additionally.*

To test the three stated predictions I use a unique dataset of Swiss Air Force pilots' engagements and subjective performance evaluations.

3.4. Empirical Analysis

3.4.1 Data

I collected my data by combining individual datasets: Swiss Air Force pilots have been required to state the nature of what they do on a daily basis. Each evening they complete one column in an individual excel-sheet by declaring the number of hours they have worked on different tasks. Since these data are purely self-declarative for military statistical use, there is no reason to assume that the reports are biased. The data were furthermore not available to direct superiors! This is a very lucky situation for a researcher as I consider the amount of hours spent on extra duties as truly unobserved by the direct superior. I however am in the possession of all the numbers and can therefore test my predictions.

For my analysis I use the datasets starting with November 2002. The subjective performance evaluation system that had been introduced two years earlier only then became wage effective.[87] Due to the introduction of a new IT-solution that had some problems in the beginning, a lot of data were lost in 2005. Consequently, I limit my observation period by October 2004 with the end of an evaluation period. After transforming the data of the available excel-files into the statistics program Stata, I generated missing values if pilots claimed to work more than 24 or less than 0 hours a day, which presumably had to be typing errors. I furthermore deleted pilots who only formally filled out their excel-sheets with a daily 8.5 hours.[88] Moreover for certain duties I had to generate hourly figures from daily data. For example, a singular day-off was generated as 8.5 hours free time, reflecting the standard daily

[87] It actually became wage effective by January 2003. I consider November 2002 the more interesting date as the employees then had received their evaluations and therefore could already be happy or disappointed about their next year's salary.

[88] According to tests this removal does not influence the later findings.

working requirement of similar jobs in the government.[89] As the pilots had to submit a monthly excel-sheet to headquarters, on a few occasions a whole month was missing.[90] To correct for this circumstance I collapsed the daily data into monthly data and generated missing values for months with sums of zero working and off-duty time. I then generated a yearly mean for monthly working hours for each evaluation period.[91] This means my first period starts by November 2002 and ends by October 2003 and my second period covers November 2003 until October 2004. Next I combined all normal-duty and all extra-duty variables into two single variables. The complete list of duties is provided in Table 21 in the appendix. A test for correct affiliation will be discussed in the results section of this chapter. The descriptive statistics of all variables are displayed in Table 6. For the later analysis I use the 96 pilots for whom I have a full set of observations available and also data on their personal LOBE-evaluations for the periods 2001, 2002, 2003 and 2004.[92] According to the shape of these LOBE data displayed in Table 6 I believe that corner-solutions should not pose a problem for the further analysis, as evaluations of 1 or 5 have never been awarded. For the later analysis, I generated three different LOBE variables: As shown in Table 7, I first combined the topical working hours with the LOBE-evaluation received at the end of such a period and called it LOBE. Then I combined the topical data with the LOBE-evaluation received in the period *t-1* (LOBElag). I finally combined the topical working data with the difference of the last two evaluations and called it LOBElagdelta.[93] The descriptive statistics

[89] I also replaced a day by the individual average of working hours. This did not affect any of the later results.

[90] I furthermore found pilots who did not take great care in filing the excel-forms and most probably forgot to fill in half of the month. I therefore also turned all those into missing values who declared in less than 75% of the time what they were doing (including days off or sick leave).

[91] This helps me to state correct (conservative) standard errors in the empirical analysis as I only have one evaluation per year as opposed to per month or day.

[92] During the critical period approximately 200 pilots have been employed by the Swiss Armed Forces. Some of the pilots have however been working in other branches than the Air Force, some have been retired and some only hired in the critical period. At the end of 2005, the former command of all professional pilots has been suspended and all pilots have been reassigned to different airbases. Before that happened, I was provided with all the available data by the commander himself. Today it is very difficult to track the reasons of why specific data is missing. It seems however very likely that it is missing on a random basis and should therefore not affect the representative quality of the later findings.

[93] For further details see Table 7.

in Table 6 show that I should have enough variance in my data to perform reasonable regression-analyses.[94]

Table 6: Descriptive Statistics.

Variable	Year	Obs	Mean	Std. Dev.	Min	Max	Description
Extra Duty	03	96	42.664	31.290	0.000	155.917	The average of monthly extra hours worked over one qualifciation period
	04	96	52.404	31.236	6.000	167.000	
Normal Duty	03	96	107.228	29.981	33.889	160.542	The average of monthly nornal hours worked over one qualifciation period
	04	96	95.967	29.640	5.400	145.200	
Delta Extra Duty	04	96	9.740	22.113	-51.083	93.677	Difference between Extra Duty 04 and 03
Delta Normal Duty	04	96	-11.260	26.579	-86.173	74.042	Difference between Normal Duty 04 and 03
LOBE value	01	96	3.408	0.376	2.700	4.200	LOBE-Evaluation at the end of a qualification period
	02	96	3.299	0.283	2.800	4.100	
	03	96	3.319	0.267	3.000	4.000	
	04	96	3.341	0.248	2.600	4.100	
LOBElag value	02	96	3.408	0.376	2.700	4.200	LOBE-Evaluation of the last period (t-1) (LOBElag02=LOBE01)
	03	96	3.299	0.283	2.800	4.100	
	04	96	3.319	0.267	3.000	4.000	
LOBElagdelta	03	96	-0.109	0.314	-1.000	0.500	Difference of last two LOBE Eval. (LOBElagdelta03 = LOBE02-LOBE01)
	04	96	0.020	0.180	-0.400	0.500	
Wage bracket	03	96	23.792	2.271	17.000	28.000	The wage bracket of a worker
	04	96	24.667	1.185	24.000	29.000	

Notes:
Source: own calculations based on data provided by the Swiss Air Force.

Table 7: Further Descriptions.

Period	Variable Name	Formula	Description
Evaluation Year 03 (Nov 02 - Okt 03)	Extra Duty		The average of monthly hours worked on extra duties between November 2002 and October 2003
	Normal Duty		The average of monthly hours worked on normal duties between November 2002 and October 2003
	LOBE	LOBE03	The LOBE-evaluation received by October 2003
	LOBElag	LOBE02	The LOBE-evaluation received by October 2002
	LOBElagdelta	LOBE02-LOBE01	The Difference between the LOBE-evaluation received by October 2002 and October 2001
Evaluation Year 04 (Nov 03 - Okt 04)	Extra Duty		The average of monthly extra hours worked between November 2003 and October 2004
	Normal Duty		The average of monthly hours worked on normal duties between November 2003 and October 2004
	LOBE	LOBE04	The LOBE-evaluation received by October 2004
	LOBElag	LOBE03	The LOBE-evaluation received by October 2003
	LOBElagdelta	LOBE03-LOBE02	The Difference between the LOBE-evaluation received by October 2003 and October 2002
	Delta Extra Duty	Extra04-Extra03	The Difference between the time spent on EXTRA duties in Nov 03-Oct 04 and in Nov 04 - Oct 03.
	Delta Normal Duty	Normal04-Normal03	The Difference between the time spent on NORMAL duties in Nov 03-Oct 04 and in Nov 04 - Oct 03.

Notes:
Source: own describtions based on data provided by the Swiss Air Force.

[94] As I am going to use fixed effects regression techniques it is critical that my data bears inter temporal differences. This should be the case as later shown in Figure 12 - Figure 14. Out of the 96 pilots I have only 35 with zero differences in LOBElag between 2004 and 2003 and only 17 with zero differences in LOBElagdelta between 2004 and 2003. Finally there are no pilots that performed in 2003 and 2004 the same amount of extra duties.

3.4.2 Analysis

Before testing the propositions I have to assess whether performing higher "unobserved effort levels", i.e. working more extra duty time, does not indeed improve the subjective performance evaluation of a pilot. If it would, extra duty time could not be regarded as "unobserved effort". To test whether the amount of performed duty influences the subjective performance evaluation, particularly extra duty, I regress the evaluation on the average monthly hours spent for normal and extra duties. The formal statistical analysis, corrected for clustering, can be found in the first column of Table 8.

Table 8: Regressions of the effects of different duties on the later subjective performance evaluation LOBE.

LOBE	OLS Full Set of Observations		Fixed Effects Full Set of Observations		OLS Restricted Set of Observations		Fixed Effects Restricted Set of Observations	
Normal duty	-0.001		0.000		0.001		0.000	
	(0.001)		(0.001)		(0.001)		(0.001)	
Extra duty	0.002	**	-0.001		0.003	***	-0.002	
	(0.001)		(0.001)		(0.001)		(0.001)	
Nov03-Okt04	-0.033		0.036	*	-0.006		0.044	*
	(0.024)		(0.021)		(0.028)		(0.026)	
const	0.002	**	3.306	***	3.114	***	3.262	***
	(0.001)		(0.132)		(0.149)		(0.132)	
adj. R^2	0.1775		0.7004		0.1408		0.5953	
(within)			0.0394				0.0490	
n	271		271		192		192	
clusters	150		150		96		96	

Notes:
Dependent variable: Qualification Rating
Coefficient of OLS and Fixed-Effects regression
(Robust standard errors in parentheses, OLS adjusted for clustering)
Level of sign.: ***:p<=0.01,**: 0.01<p<=0.05, *:p<=0.10
Source: own calculations based on Swiss Air Force data.

I also include a dummy for the second evaluation period from November 2003 until October 2004 to prevent external influences from biasing my regression. The base-group is the evaluation period from November 2002 until October 2003. According to these results working on average one more hour per month on extra duties actually increases the LOBE-evaluation of a pilot by 0.002 points. This preliminary result seems to overthrow the basis for my propositions. It does not however make use of the panel nature of my dataset. While I want to know whether persons receive a better subjective performance evaluation because they work more extra duty time, the results could as well mean that persons working more

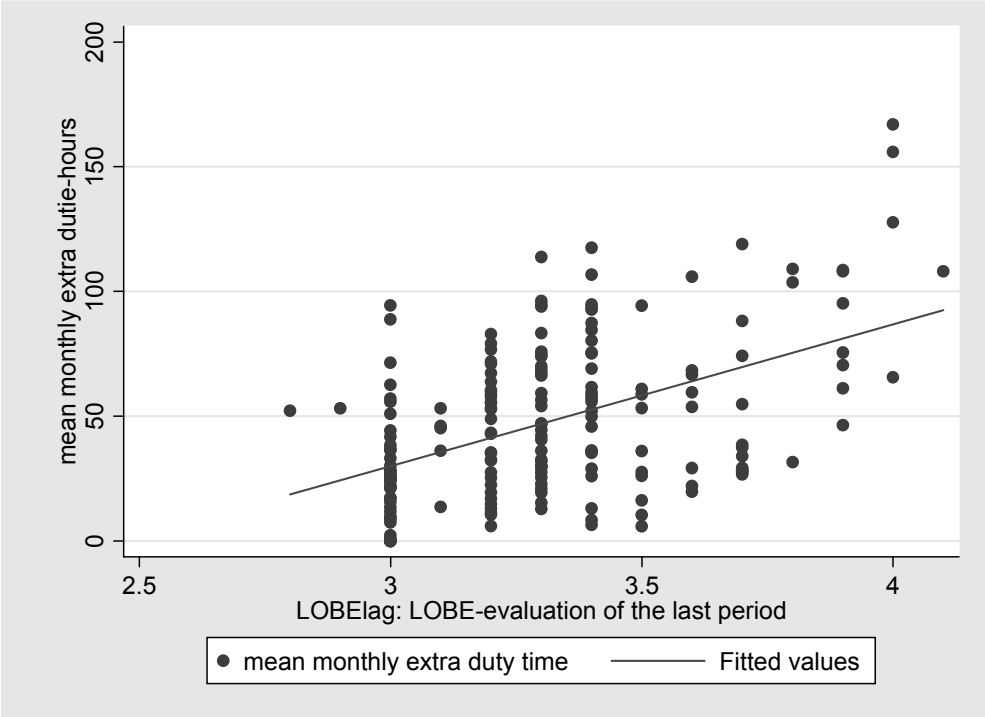
extra duty time are the same that receive on average a better subjective performance evaluation, but due to other reasons. It is consequently critical to the analysis to make inter-temporal regressions. Such a more stringent statistical analysis can be found in the second column of Table 8 where I proceed with a fixed-effect regression. I again include a dummy for the second evaluation period to allow for annual changes. The effect of extra-duty-hours on the LOBE-evaluation at the end of a period becomes statistically insignificant and even changes its direction (see column two in Table 8).[95] Hence I cannot reject the hypothesis that the hours spent on extra duties do not influence the LOBE-evaluation of an individual pilot. As for the further analyses I am going to use a sub-sample of the pilots due to data-availability, I re-run the same regressions for the sub-sample and display the results in the third and fourth column of Table 8. The results are unchanged.[96] Therefore the underlying condition that the unobserved effort level does not increase the subjective performance evaluation seems to be met in the case of extra duties. I postpone any further interpretation of the results to the end of this section.

I now proceed and assess my first proposition. A graphical analysis of the relationship between the received subjective performance evaluation and the amount of time spent for extra duties thereafter is presented in Figure 12. On the x-axis I display the LOBE-evaluation of the last year (LOBElag) and on the y-axis the mean monthly amount of performed extra duties this year. The fitted values suggest a linear combination of the two effects. So I regress the effect of last year's result of the subjective performance evaluation on the hours a pilot works on extra duties per month and correct for salary effects caused by management positions by including the wage bracket of the pilot. The results of a standard OLS regression with cluster analysis and the compulsory dummy-variable for the second evaluation period are displayed in column one of Table 9. According to this result it seems as if pilots exert many extra duties if they receive a good subjective performance evaluation. But the result may be biased in a similar way as the OLS regression of the effect of extra duty on the LOBE-evaluation before. To make sure that only a change in the LOBE-evaluation is relevant for my analysis, I proceed by using fixed-effects regression techniques.

[95] As I report the full R^2 of the regression, which is fairly high due to explanative power of the fixed effects, I also include the "within R^2" which shows that the remaining estimators are not able to explain much of the variance in the data.

[96] To further test this issue I run the same two regressions with only those pilots who had different LOBE-values for the years 2003 and 2004. The results remain almost the same. Even the significance of the estimator for extra-duty in the OLS regression only changes by a mere 0.1%.

Figure 12: Interdependence of the last subjective performance evaluation and unobserved effort level.



Notes: Source: own calculations based on data by Swiss Air Force.

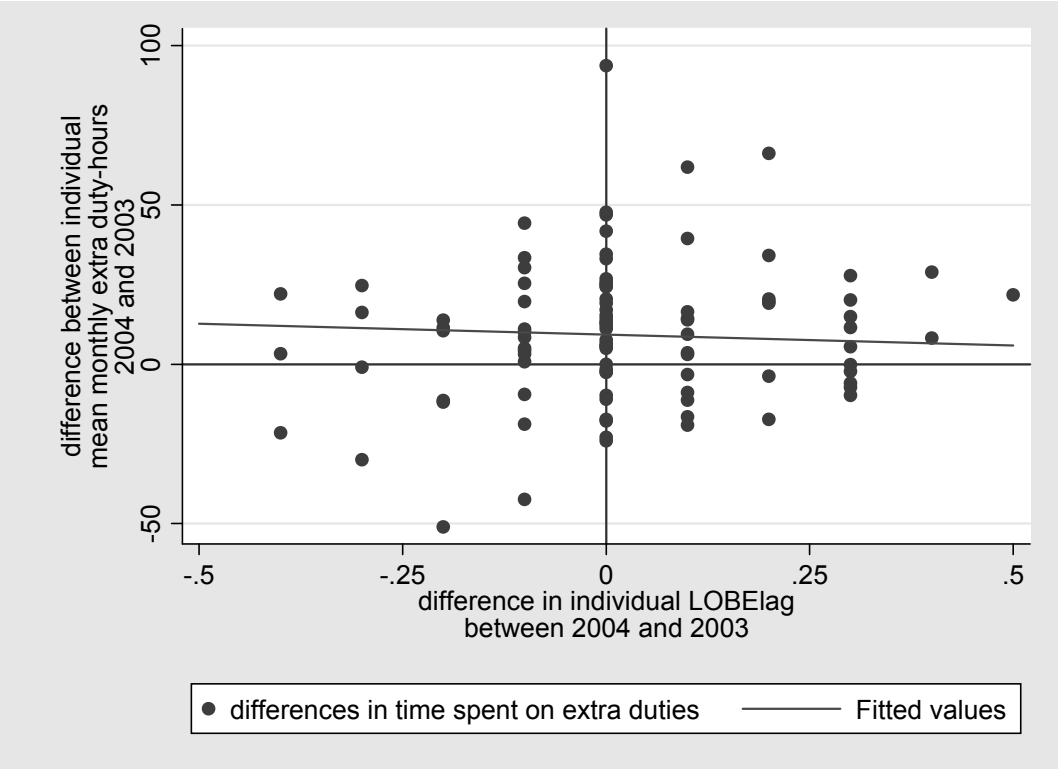
Table 9: Regression of the effect of the last evaluation LOBE on the mean monthly extra duty time.

Extra duty monthly hours	OLS	Fixed Effects	OLS	Fixed Effects	Fixed Effects standard errors adusted for clusters	Fixed Effects weighted for 1/var & month/12
LOBElag	33.569 ***	16.512				
	(9.169)	(12.409)				
LOBElagdelta			10.817 *	14.004 **	14.004 *	7.189 *
			(5.981)	(5.577)	(7.937)	(4.284)
Wage bracket	7.467 ***	2.081 *	9.860 ***	2.222 **	2.222 **	1.729
	(1.300)	(1.136)	(1.501)	(1.110)	(0.930)	(1.070)
Nov03-Okt04	2.542	7.592 ***	-0.280	5.994 **	5.994	5.023 ***
	(2.602)	(2.442)	(2.914)	(2.493)	(3.955)	(1.790)
const	-245.720 ***	-61.324	-190.748 ***	-8.667	-8.667	-24.329
	(35.056)	(48.584)	(35.242)	(26.448)	(21.238)	(21.182)
adj. R^2	0.4069	0.7626	0.3474	0.7735	0.7735	0.9062
n	192	192	192	192	192	185
clusters	96	96	96	96	96	96

Notes:
Dependent variable: yearly mean of monthly extra hours
Coefficient of OLS and Fixed-Effects regression
(Robust standard errors in parentheses, OLS adjusted for clustering)
Level of sign.: ***:p<=0.01,**: 0.01<p<=0.05, *:p<=0.10
Source: own calculations based on Swiss Air Force data.

The graphical analysis of this method is easy: I just compare the effects of the differences in one variable to the differences in the other variable. Figure 13 does this and displays the difference between the last received LOBE-evaluation and the one before that on the x-axis and the difference between the amounts of performed extra-duty this and last year on the y-axis. In this way I can assess whether a change in the evaluation influences the pilot to change his behaviour. Figure 13 draws a less clear picture than Figure 12.

Figure 13: Interdependence of the difference between the last two subjective performance evaluations and the difference between the actual and the last unobserved effort levels.



Notes: Source: own calculations based on data by Swiss Air Force.

The fitted values do not show a positive linear relationship between the variables on the two axes. This graphical analysis is confirmed by an analogue fixed effects regression with a periodical dummy in column two of Table 9.[97] Even though the coefficient points in the right direction, according to large standard errors I must reject the hypothesis that a pilot executes

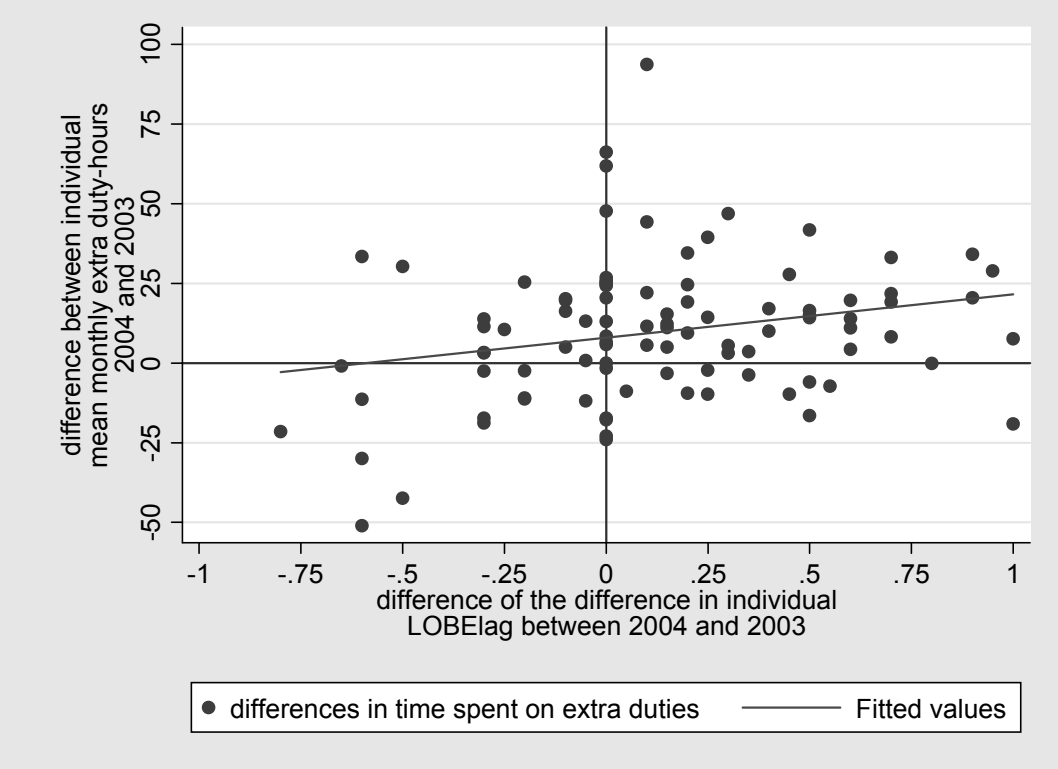
[97] To further test this issue and render the OLS and the Fixed-Effects regressions comparable, I rerun the OLS regression of the first column of Table 9 with only those pilots who had different LOBElag-values for the years 2003 and 2004. This time the results of the OLS regression change slightly, but *extra-duty* remains significant at the 5% level. Hence it is not the sample selection but truly the method that change the results of the two regressions.

more extra duties if he receives a better subjective performance evaluation, which would have reflected the left side of Figure 10 on page 47.[98] Only the wage bracket remains relevant, as pilots who attain a higher hierarchical position and therefore enter a higher salary-class seem to work more on extra duties afterwards.

- ***Result 1:*** *The time a pilot spends on extra-duties is not a function of the result of his most recent subjective performance evaluation.*

To further understand these issues I proceed by analysing my second proposition. To test whether the difference between the last two LOBE evaluations, namely LOBElagdelta, influences the amount of extra duties a pilot performs, I start by a graphical analysis: To make intra-personal comparisons only, I apply again in Figure 14 my "graphical fixed effects regression".

Figure 14: Interdependence of the difference of the difference of the last two subjective performance evaluations and the difference of the difference between the actual and the last unobserved effort levels.



Notes: Source: own calculations based on data by Swiss Air Force.

[98] The difference in the labelling of the x and y axis of Figure 13 in reference to Figure 10 on page 47 are caused by the graphical analysis of the fixed effects method.

The y-axis is again the difference between the amounts of extra duties performed this year and the year before. On the x-axis I however use the difference between the LOBElagdelta of this year and the one of last year.[99] The fitted values in Figure 14 hint at a positive linear combination of the two axes. This means that an increase in LOBElagdelta results in an increase in the performed extra duties. To more sophisticatedly test this picture I run a OLS and a fixed effects regression with the average time spent on extra duties as the dependent variable and the difference of the last two subjective performance evaluations (LOBElagdelta) as the explanatory variable. The third and the fourth column of Table 9 show the results. This time the proposition seems to be confirmed for both cases. I find a statistically significant effect of the difference of the last two subjective performance evaluations on extra duties. According to these regressions I cannot reject the null-hypothesis that pilots perform more extra duties if they receive a better subjective performance evaluation than before. In the case of the fixed effects regression, the statistical significance level is 1.4%. These results are astonishing all the more as the first proposition has not been supported. Thus they solely encourage the view expressed in the right panels of Figure 10 on page 47.

- ***Result 2:*** *The time a pilot spends on extra-duties is a function of the difference between his most recent subjective performance evaluation and the previous one.*

Due to the importance of this result I run three further tests: First I adjust the standard errors in the regression for clustering on the individuals.[100] According to the fifth column of Table 9 my findings are only slightly less significant. Second I address the construction of my dependent variable. As it is an average value I use the inverse variance of all the monthly values to weigh my regression. As furthermore some of these months are missing, I combine the weighing-factor with a factor that indicates the number of available months as a fraction of 12. This should render my regression trustworthier because observations based on big variances and observations based on few months are treated as less important than others. The sixth column in Table 9 shows that again my result remains significant at the 10% level. I lose some observations in this regression because some average values for monthly extra duties are zero and therefore without variances. As a third test I assess whether these results are critical to the affiliation of the tasks in Table 21 (see appendix). Thus I change the affiliation of every single task by deleting it from his original category and adding it to the opposite category. Then I re-run the regression of row four in Table 9. Table 10 shows the six for my

[99] This is in fact the difference between the differences between the last two LOBElag-values.

[100] With the use of adjusted standard errors, I prevent them to be too small for the case that the error-term is correlated within the observations of an individual pilot.

analysis critical tasks and their exact nature. If I change any of the tasks in the first section of the table from being part of normal duty to being part of extra duty, the statistical significance of the results of Table 9 deteriorates as shown in Table 10 to either below 5% or even below 10%. For the case of the extra duties, only one task is critical. The test is supportive as I can see that all the duties in Table 10 are clearly defined to belong to either normal flying duties or extra duties by their very nature.

Table 10: The six critical tasks for the analysis.

Regression results (Table 4)	LOBElag	14.004 ** (5.577)	original regression

removed from normal duties added to extra duties			Description
bmpsausbil~g	LOBElag	12.986 * (6.867)	professional pilot school as student
flgezgdrlt~g	LOBElag	10.702 * (6.279)	flights for third parties
flugdienst~g	LOBElag	14.025 * (7.138)	individual flight training
wau	LOBElag	8.158 (6.315)	advanced flight training
spukspugin~n	LOBElag	6.846 (6.235)	flight instructor for advanced training
flugdienstit	LOBElag	6.263 (6.639)	other individual flight training

removed from extra duties added to normal duties			Description
einsatzkos~o	LOBElag	-4.403 (4.919)	voluntary mission abroad as peace support

Notes:
Critical Coefficient for LOBElag of Fixed-Effects regression
after changing the category of every single variable.
Dependent variable: yearly mean of monthly extra hours
(Robust standard errors in parentheses).
Level of sign.: ***:p<=0.01,**: 0.01<p<=0.05, *:p<=0.10
Source: own calculations based on Swiss Air Force data.

It could only be argued that voluntary peace support missions, which actually consist of flying helicopters over the Balkans, do belong to the normal duties of a pilot. This duty is however voluntary (Swiss Government, 1995:21, Art.66 and Schindler, 1997:170) and it seems from year to year more difficult to find enough pilots to operate the helicopters the Swiss Air Force stationed in the Balkans. In 2006 it was even decided to reduce the crew ratio to a mere one crew per helicopter to cope with the problem. Hence Balkan-duty seems to me a perfect field for reciprocity.

Before making any further conclusions I test the third proposition, i.e. whether there is a substitution effect of extra for normal duties. To assess this proposition I simply estimate the effect of the differences of the mean monthly hours spent for extra duties on the differences of

the hours spent for total and normal duties. Table 11 displays the simple OLS-regression results. As I already use the differences, a yearly dummy is obsolete and an ability-bias may be excluded. Both columns in Table 11 bear significant results that logically add up to one. According to the point estimates in Table 11 a pilot who works one additional extra-duty hour, does this by working 19 minutes additionally and by substituting 41 minutes of normal duty-time.

Table 11: Regression of the difference in extra duties on the difference in total/normal duties (yearly monthly means).

OLS	Delta Total duty		Delta Normal duty	
Delta extra duty	0.321 (0.102)	***	-0.679 (0.102)	***
const	-4.647 (2.461)	*	-4.647 (2.461)	*
adj. R^2	0.0852		0.3119	
n	96		96	

Notes:
Dependent variable: Difference in yearly mean of monthly total and normal hours (04-03)
Coefficient of OLS regressions
(Robust standard errors in parentheses, OLS adjusted for clustering)
Level of sign.: ***:p<=0.01,**: 0.01<p<=0.05, *:p<=0.10
Source: own calculations based on Swiss Air Force data.

I hence cannot reject that an increase of the amount of extra duties performed does result in an increase of total time worked. But I can neither reject that an increase in the amount of extra duties performed does result in a decrease of normal duties performed. I infer that a substitution of extra for normal duties and an increase of total duties occur.

- ***Result 3:*** *If a pilot performs more extra duties, he does this by substituting some extra duties for normal duties and by performing some extra duties additionally.*

I round up this section by combining the empirical findings in one single line of argumentation: If a pilot performs more extra duties than last year, he does not necessarily get a better subjective performance evaluation according to the fixed effects estimations in Table 8. Hence the amount of time a pilot spends on extra duties may be considered as unobserved effort. If, however, he receives a better subjective performance evaluation than last year, he is very likely to work more on extra duties and therewith show reciprocity by unobserved effort. According to the fixed effects estimations in Table 9 he does so for every 0.1 points of his

evaluation for 1.4 monthly hours. If he receives a continuous high evaluation, according to the results in the second column of Table 9, he settles back to his standard unobserved effort level without exerting any additional extra duty anymore. The results of the OLS-estimations in Table 8 finally point towards the fact that pilots who always exert high unobserved effort levels seem to receive better subjective performance evaluations altogether, but not necessarily so after an especially high unobserved effort level. It may seem plausible to assume that the same pilots trigger the results of the first OLS-estimation in Table 9, even though they do not do this because they receive good subjective performance evaluations. Hence a standard omitted variable bias may exist which is avoided by using fixed effects regression techniques (Wooldridge, 2003:485).

Comparing now the power of prediction of standard incentive theory and theories of reciprocity I believe that standard incentive theory for multi-tasking has difficulties in explaining the above results. Furthermore its prediction that extra duty vanishes over time as pilots might notice the gap in the incentive system is not supported. The descriptive analysis of the difference in the yearly extra duties (Delta Extra Duty) in Table 6 shows that in the evaluation period between November 2003 and October 2004 more extra duties have been performed as in the period before. This means that the accomplishment of extra duties does not vanish over time in my data. However, theories of reciprocity are also not able to fully predict the performance of extra duties. Proposition One has not been confirmed. But theories of reciprocity are able to explain at least the dynamic side of my findings. They are able to predict who is exerting a higher unobserved effort level and who is exerting a lower unobserved effort level than last year. But initial unobserved effort levels seem to depend mainly on the general perception of a fair salary and possibly fair working conditions. Furthermore it appears as if the pilots receiving on average the better results of their subjective performance evaluations are the ones performing most extra duties, even though they do not seem to be motivated by considerations of reciprocity.

3.5. Conclusions

Over the last few years many insights have been gained into incentive systems for workers for whom output measures are easily observed (Prendergast, 1999:57). Data has however been difficult to achieve for systems with an output less easy to measure. The applied *subjective* performance evaluations have therefore only been scarcely researched. In the meantime, newly developing literature with mostly theoretical and experimental papers has shown that

many people do not care for only their individual profit maximization according to neoclassical models but also for fairness. As subjective performance evaluations are made by human beings and not by a standard piece-rate formula, it seems possible that these evaluations may be appreciated and remunerated by reciprocal employees. Merging a unique firm level dataset of operational data with performance evaluations I thus tried to find out whether changes in subjective performance evaluations induce reciprocity in the field of unobserved efforts. I further researched whether such effects last over time and if they may even resolve the problem of optimal effort allocation to unobserved tasks.

My major findings support theories of reciprocity since the subjective performance evaluation has the following effect on behaviour: If pilots receive a better subjective performance evaluation than before, they perform higher levels of unobserved effort in the next period. If they receive a worse evaluation than before, they perform lower levels of unobserved effort than in the preceding period. This behaviour does not however last. I only find results for the period right after a change in the subjective performance evaluation. This means that pilots slip back to a normal level of unobserved effort if their now adapted expectations are met, even if the expectations are at a high or low level. Hence the effects of reciprocity do not achieve a sustained success in motivating the pilots to perform unobserved tasks. The effects are short lived. All the same, time spent on unobserved tasks does not vanish over the years. It seems thus probable that a pilot chooses a certain level of unobserved effort that he perceives as fair and thereafter only changes it for a short time if his expectations are not met or exceeded.

These findings shed light on three issues: First it appears as if standard incentive theory is not able to explain the chosen allocation of effort in an environment with subjective performance evaluations and multiple tasks of which parts are unobservable. Even if the incentive system has a gap, appropriate working conditions and the fairness of the workers may ensure the performance of unobserved tasks at a certain level. Second, the subjectivity of subjective performance evaluations seems to matter. If the evaluation is changed, the worker is either thrilled or disappointed. He then changes his chosen unobserved effort level according to the degree to which his expectations are exceeded. Hence, reciprocity indeed plays a role. Third it appears as if expectations adapt very quickly to new situations. These expectations not only cover absolute salary but even salary changes. Thus, even high salary rises might become expected and for this reason normal.

If accordingly workers in complex professions should be motivated by reciprocity to provide effort beyond normal levels, an exponential wage path would have to be applied. And even then, as soon as their expectations adapt, the additional effort levels might vanish.[101] Consequently the use of wage effective incentive systems should be reconsidered in professions where effort is difficult to measure.[102]

Whether it at least helps to employ workers that have strong positive reciprocal preferences will show the next chapter. I will research whether the experimentally found strong, though still varying, positive reciprocity of the pilots in chapter two predicts their real reciprocal behaviour towards the Air Force.

[101] And what about bonus payments? Bellmare and Shearer (2007:17) find for example that tree planters reciprocate a gift given on one day with extra high effort levels. These high effort levels than vanish over the next few days without a tendency to go below the original level. This means that no other gift has been expected. In the long run, it seems however unlikely that the mere fact that bonus payments are possible in a certain firm does not raise expectations. Therefore also bonus payments are not a lasting solution to the problem.

[102] Already Prendergast (1999:30) reports that many managers disconnect evaluations from wage altogether to avoid the endangerment of the good prolonged relationship of employers and employees.

CHAPTER 4:

Reciprocity at Work

Evidence from matched experimental and personnel data.

4.1. Introduction

It has been long recognized that social preferences, in particular that of reciprocity, can have profound implications for internal labour market relations (Akerlof and Yellen, 1986).[103] Indeed, evidence from laboratory experiments demonstrates large effects of social preferences on the behaviour of its participants.[104] Nevertheless, many economists question its applicability to, or its significance for, the real-life workplace (e.g. Harrison and List, 2004:1009f; Levitt and List, 2007; Gneezy and List, 2006). Either they claim that the subjects in experiments are mostly students, who do not necessarily represent accurately the behaviour

[103] There is a growing literature of many different theoretical approaches that all support the idea that people not only act to enhance their own material well-being but also use costly actions to equalize the distribution of final payoffs (Akerlof, 1982; Akerlof and Yellen, 1988 and 1990; Levine, 1998; Fehr and Schmidt, 1999; Falk and Fischbacher, 2000; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Camerer and Fehr, 2006). It is not of my concern whether it is just inequality aversion or indeed the urge to reward and revenge as I will not distinguish the different theories but take it as given that reciprocal actions are proposed.

[104] See e.g. Fehr, Kirchsteiger and Riedl (1993), Berg, Dickhaut and McCabe (1995), Fehr, Gachter and Kirchsteiger (1997), Fehr and Falk (1999), Fehr and Gachter (2000b), Falk and Gachter (2002), Andreoni, Harbaugh and Vesterlund (2003), Fehr and Schmidt (2004), Fehr and List (2004), Charness (2004) and Gachter, Kessler and Koenigstein (2006).

of workers (Fehr and List, 2004; Carpenter, Burks and Verhoogen, 2005) [105] or that the settings do not resemble sufficiently enough real life situations (Kube, Maréchal and Puppe, 2006). Dohmen et al. (2006:11), for instance, find that students – the typical subjects in experiments - are in these settings less reciprocal than the general population. And Lazear, Malmendier and Weber (2005) also found experimental results to be biased because real people self-select into different settings.[106] But how is it possible to determine whether reciprocity plays an actual role in real labour markets? Up to now, there have been three other commonly used avenues of research: Surveys, field experiments and the analysis of real-life data, all of which have their own advantages and disadvantages. Surveys, as a common alternative to experiments are not only based on students, but also lack the other problem of credibility: A person might claim one thing and do another. Field experiments seem to be a valuable alternative to construct real life situations (List, 2007). These three studies manipulate the 'kindness of the wage' (Gneezy and List, 2006; Kube, Maréchal and Puppe, 2006; Bellmare and Shearer, 2007).[107] The design of these experiments is set up to examine the "pure" form of reciprocity in one-shot interactions without any scope of a long-run relationship. Removing concerns for reputation is important, otherwise even strictly selfish individuals may have an incentive to mimic reciprocal behaviour to reduce the danger of being replaced (Brown, Falk and Fehr, 2004). Consistent with the experimental evidence (e.g., Falk and Gachter, 2003), these papers find only moderate effects of reciprocity on behaviour. Real life data finally, if possible to gather at all, are always influenced by many factors researchers are not able to control nor even grasp. Hence it is difficult to tell whether the observed phenomena are really based on individual reciprocity or whether other environmental factors in addition to reputational considerations trigger the behaviour.[108]

---

[105] The most critic received experiments with students majoring in economics as they have been shown to behave differently compared to other students (Carter and Irons, 1991; Frank, Gilovich and Regan, 1993). Furthermore Bellmare and Kroeger (2005) compared students' decisions in an investment game with those of subjects of a representative panel in Denmark. They found that students' decisions are only to a limited extent representative for the population (p.8).

[106] The same line of argument follow Eriksson and Villeval (2004), Dohmen and Falk (2006) and Eriksson, Teyssier and Villeval (2006).

[107] Gneezy and List (2006) and Kube, Maréchal and Puppe (2006) employed students to computerize the holdings of small libraries and (only Gneezy and List, 2006) to take part in a door-to-door fundraising drive. Bellmare and Sherer (2007) granted tree planters a salary increase for only one day.

[108] Some studies are Katz, Kochan and Gobeille (1983), Cappelli and Chauvin (1991), Kleiner, Leonard and Pilarski (2002), Bartel et al. (2004), Mas (2004), Krueger and Mas (2004) and Mas (2006).

In this chapter I will try to address these problems by treading a new path of research: I will match experimental findings of reciprocity with real-life data.[109] I use the measurements of reciprocity obtained by laboratory experiments among employees of a large organization (see chapter two) and compare individual values for reciprocity to the apparent reciprocal behaviour of the same employees over a two-year period in a real labour market situation (see chapter three). This provides me with an opportunity to examine whether the behaviour observed in the workplace is indeed based on reciprocity as measured in experiments. Thus, my strategy is to measure reciprocity in a one-shot experiment, where reputation plays no role, and to examine whether this predicts the differences in behaviour within the workplace that cannot be explained by reputational concerns. Therefore, rather than manipulating the environment in the workplace, I use naturally occurring changes in the workplace and examine whether individual differences in reciprocity predict the response to these changes.[110]

For the experiments, I use the so-called "moon lighting game" (Abbink, Irlenbusch and Renner, 2000). It puts the subject into a position where he or she has to choose whether to reward or punish a friendly and an unfriendly action. Consequently I am able to measure both, positive and negative reciprocity in the experiments.

To measure the behaviour of my experimental subjects in the workplace, I make use of its specific setting. The subjects, like myself, are pilots of the Swiss Air Force, a multi-task environment with subjective performance evaluations. These evaluations directly affect the salary increase of the pilots. However, supervisors tend not to observe all the dimensions of their pilot's efforts. While supervisors base evaluations mostly on subjectivity, some aspects of work effort, in particular the time pilots spend on extra duties, like for example, the

[109] Outside the labour market context, several other papers related behaviour in experiments to behaviour in real-life situations: Fehr et al. (2003) conducted a survey and included experimental questions to assess which survey questions predict the behaviour in the experimental questions. They found that trustworthiness in their gift exchange game is not predicted by the claimed trusting behaviour in the past of the individuals. Carter and Castillo (2002) researched whether the results of trust games may serve as a predictor for the material well being of poor people living in South Africa and found positive results for urban areas but not for the rural ones. Bettinger and Slonim (2005) conducted dictator games in schools and researched whether the winning of a scholarship changes the behaviour of children and parents but found only results for their behaviour towards charitable organisations but not towards fellow students.

[110] This chapter is most closely related to Karlan (2004), who compared laboratory trust games with field data in Peru. He found that the more trustworthy people in experiments had a higher probability to pay back their loans in a real group lending microfinance program.

evaluation of a new system, go largely unnoticed. This means that this particular effort can not be motivated by considerations of evaluation or reputation, indicating all the more that reciprocity might be a relevant factor: Due to the subjective nature of the evaluations, a worker might reciprocate the announcement of a positive or negative change in his subjective performance evaluation with an effort that does not affect his salary.

By using a unique dataset that allows me to track such activities, I find that pilots vary their effort level on extra duties in response to changes in their performance evaluations. Surprisingly good evaluations lead to higher effort levels on these unrewarded tasks. This finding alone, however, may not only be attributed to reciprocity. It can also be construed that pilots with good evaluations are asked to perform these tasks. To attribute the behaviour to reciprocity, I examine whether the individual reciprocity measured in the experiment predicts responses to changes in the performance evaluations. And indeed, I find that those pilots, who demonstrate more strongly positive reciprocity, increase their effort levels more after a surprisingly good evaluation, while pilots who are more negatively reciprocal reduce their effort levels proportionately more after a worse evaluation. Due to the small sample (roughly 60 pilots observed over a four year period), my estimated effects are not very precise. However, the variation in behaviour in the laboratory experiment is nevertheless sufficient to detect significant differences in responses to good and bad evaluations.

The rest of this chapter is structured as follows: In section two I provide the institutional background of the specific workforce, section three explains the experiment and section four renders the predictions. The empirical analysis and the results are discussed in section five. Finally, section six is the summary and conclusion.

4.2. Institutional Background

I conducted my experiments (see chapter two) and empirical analysis (see chapter three) in an environment that has some special features. It is therefore worthwhile to give the reader the necessary information to be able to appreciate the later results.[111]

The workers I am looking at are all Swiss Air Force pilots. Two special features of their situation render me the unique possibility to find reciprocity in their real life situation. First, pilots earn their salary according to a subjective performance evaluation of which its subjective characteristics might give rise to considerations of reciprocity. Second, institutional

[111] For a detailed discussion of the environmental background see chapter two and three.

arrangements within the Swiss Air Force are such that direct superiors do not have data available on an important dimension of an effort, namely how much time the pilots spend on extra duties besides their normal flight activities. This renders an opportunity for reciprocal activities.

Let me begin with the first issue: The pilots earn their salary according to three schemes. The first scheme concerns a fixed compensation received in accordance with compensating wage-differentials (Borjas, 2002, 201-225) due to their increased accident probability in military aviation (Federal Department of Defence, 2003b:1272, Art.2).[112] The second scheme concerns the so-called wage bracket. The wage bracket defines the maximum salary a pilot in a certain managerial position may reach. It hence defines the range of the salary. If a pilot gets promoted from one wage bracket into another, he gets a once-only salary rise of 5%. The last and central scheme defines salary-increase and bonus payments for pilots. It is the subjective performance evaluation system called LOBE. This acronym stands for "Lohnrelevante Beurteilung" and translates as "wage-relevant evaluation" (or literally “to praise”). Every year in November the superior defines individual criteria, upon which the pilot is measured throughout the following year. Examples for these criteria include “shows a strong willingness to fulfil his tasks” or else “uses wisely his freedom of action”. The superior then judges to what degree the pilot fulfils these criteria. At the end of that year, in October, the pilot receives the results of this subjective performance evaluation on a scale of 1-5. If he receives a result of 1 he does not get any salary raise and may fear layoff. If he receives a result of 5, he will get a 6% wage-increment for the next period. If he already reached the maximum salary of his wage bracket he can get as much as a 12% bonus payment. There are however budget constraints on how many pilots get good evaluations. It is defined that a maximum of 18% of the workers might get a bonus payment above 4.4% (Swiss Government, 2001:21, Art.49).[113] It is important to note that the measurement of the defined criteria is of a purely subjective nature, as superiors have no access to the statistical data about, for example, the amount of hours a pilot actually works.[114] And so pleasing a superior becomes very

[112] This accident probability is not to be confused with the economic risk (see e.g. Bloom and Milkovich, 1998).

[113] A tournament based on this forced distribution is unlikely because the pilots are evaluated upon different criteria.

[114] This astonishing fact is caused by the different interests of the people who collected the data about the time spent on different tasks by the pilots and the ones responsible for the subjective performance evaluations of the pilots. Hence the information I was able to gather has never been passed on to the direct superiors of the pilots because it has never been edited in this way.

difficult if personal disliking of one another exists.[115] For this reason, and because it is wage-relevant, I use the subjective performance evaluation as a proxy for friendly and unfriendly actions by superiors, which might give rise to considerations of reciprocal friendly and unfriendly actions by the pilots.

The second issue now renders the pilot the perfect opportunity to exert reciprocity by acting friendly or unfriendly himself: Although the subjective performance evaluation system is meant to consider all of the pilot's tasks,[116] gaps do exist: I will show later that the amount of extra duties performed by a pilot does not directly influence his subjective performance evaluation.[117] This means that I not only believe, like Gibbons (1998:121), that complex effort is unverifiable, but also show that changes in effort levels on extra duties are unobserved by superiors. As this opens a free field for reciprocal activity without detrimental effects on reputation, I gathered this critical data, which was possible because my direct superior was Commander of all Swiss Air Force pilots.

The described situation may hold a general relevance: There is in every job some unobserved effort level, where normally nobody is able to measure it. According to standard incentive theory, workers would avoid doing extra duties if their accomplishment is not wage effective (Prendergast, 1999: 21-33). In the particular case of the Air Force, the execution of such duties would be difficult due to the lack of volunteers available, since pilots, by nature, like to fly much more.[118] However, if extra duties, which are in themselves complements to normal duties, are not performed properly, the overall performance of the Air Force suffers.[119] Hence

[115] For these cases the subjectivity of the evaluation is also emphasized by Schettgen (1996:261). For further points see also Becker (1999:369) and Becker (1998: 241).

[116] Baker (2000:420) suggested the use of subjective performance evaluations to exactly solve the multitasking problems of today's complex jobs.

[117] With extra duties, I consider all the duties a normal pilot does not do. A non complete list of extra duties would contain representative work, the organisation of operations, the work for superiors or other organisations as representatives of the Air Force, the publication of new manuals, the re- and evaluation of equipment, work in the field of human resources and IT, pre-screening or flight safety activities, missions abroad and other special projects. For a complete list see Table 21 in the appendix.

[118] This seems reasonable because these pilots underwent a prolonged screening to become a military aviator. Thus pilots must be highly motivated to fly, i.e. to primarily exert normal duties.

[119] This information is important because Hennig-Schmidt, Rockenbach and Sadrieh (2005) found in laboratory experiments that workers show positive reciprocity if they know about the employer's surplus.

voluntary performance of extra duties may be considered as friendly behaviour of pilots by which means they may express their reciprocity.[120]

Aside from the possibility to measure reciprocity in their real life situation, I must bear in mind that I am not conducting research within an average workforce. Like Lazear, Malmendier and Weber (2005:1f) suggested, people self-select into specific labour market settings. In my case a prolonged screening by the employer augments the self-selection of employees. Only 1-2% of the applicants actually become Swiss Air Force pilots. This reflects not only the supply and demand for this job but also the fraction of applicants that is able to show the required qualities to become a military pilot. This means the Air Force makes every effort to screen applicants correctly. The evaluation consists of many phases, including written exams, flying small and large training aircraft and also a computerized test for teamwork abilities. This teamwork ability plays a central role in a pilot's every day life as his survival may depend on good teamwork in combat or emergency situations. This enhanced ability may influence my empirical findings, as pilots seem to be more positive and less negative reciprocal than other people (see chapter two).

4.3. Measuring Positive and Negative Reciprocity

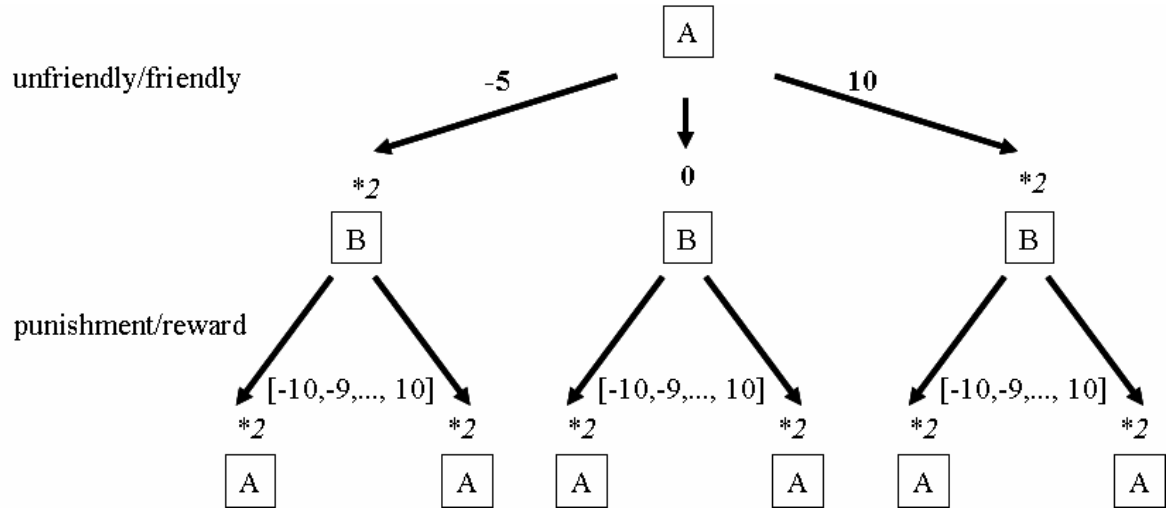
To assess positive and negative reciprocity I used in chapter two the "moon lighting game" (Abbink, Irlenbusch and Renner, 2000). I explain the course of the game according to its visualization in Figure 15.[121] In the beginning Player A and B each receive an initial endowment of 20 tokens. In the first step of the game Player A has to decide whether to be friendly, neutral or unfriendly. If he decides on being friendly he passes 10 tokens to Player B, if he is unfriendly he takes five tokens from Player B and if he is neutral, nothing happens. To render the friendly action efficient and the unfriendly action inefficient, the experimenter doubles the impact on B's payoff. That means that Player B either gains 20 or 0 tokens, or he looses 10 tokens contingent on the friendliness of Player A. The second stage is now of central interest to us: Player B has to decide how to react. He may either reward or punish Player A. Unlike Player A, he can only spend tokens to do this. Player B might use up to 10 tokens for this action. If he rewards Player A, the experimenter doubles the tokens again, so

[120] This idea is further supported by the fact that every superior is willing to allow a pilot to perform extra instead of normal duties.

[121] Figure 15 equals Figure 1. Because each chapter of this work represents a complete independent research article, I explain the experimental design of chapter two again.

Player A may receive up to 20 tokens. If Player B punishes Player A, the impact is also doubled. That means 10 tokens spent by Player B for punishment result in 20 tokens removed from Player A by the experimenter. Player B does not however receive these tokens.

Figure 15: Experimental Design: Player A's choices vs. Player B's strategic options.



Examples: *- A takes 5, so B looses 10, then B punishes -10, he has to pay 10, so A looses 20.*
- A gives 10, so B gets 20, then B rewards +10, he has to pay 10, so A wins 20.

The sub game-perfect equilibrium with strictly selfish preferences is the following: Player A foresees that Player B maximizes his payoff by not spending any tokens. Hence Player A maximizes his own payoff by taking all the possible tokens from Player B. Thus Player A ends up with his endowed 20 tokens plus the five he got from Player B. Player B can only hold on to the 10 tokens Player A could not take from him. He has no possibility to augment his own payoff. If however Player B has no strictly selfish preferences, his reactions may serve as a proxy for reciprocity. He might show positive reciprocity if he spends tokens to reward friendly actions of Player A and negative reciprocity if he spends tokens to punish unfriendly actions of Player A.[122]

I gained my experimental data on the occasion of a workshop for all Swiss Air Force pilots held in December 2005 (see chapter two). Thanks to the compulsory nature of the workshop, participation bias seems unlikely.[123] To assess for all pilots their reciprocity parameters, I

[122] I do not analyse the actions of player A, as they are not of interest for the further analysis. For more details see chapter two.

[123] Only few pilots were unavailable due to important missions that could not be cancelled.

paired the pilots with outsiders, namely students as Player A.[124] Furthermore I used the strategy method to assess positive and negative reciprocity. This means that Player B had to make decisions for all possible actions of Player A before he saw the actual decision. This had also the advantage that spontaneous emotional responses could be avoided (see e.g. Loewenstein, 2005). Like this the situation resembled much more average behaviour in a prolonged relationship between employees and employers.

On the actual day in December 2005, pilots were seated in a big hangar and were told that they could earn real money by making decisions. First they received the instructions with examples, and then the decision sheets were passed out. Only after the pilots had returned these papers, they were issued a closed envelope with the decision of Player A inside. Payments were distributed at the end of the day.

4.4. Propositions

In this section I am going to make propositions about the predictive power of experimental data for the behaviour in real life situations. To be able to do this, I start by predicting the behaviour of the pilots in their specific situation.

According to the above discussion, I regard the subjective performance evaluation as the initial friendly or unfriendly action by the superior. To assess the friendliness of this action, I make the assumption that a pilot normally expects the same subjective performance evaluation as last year. This might not always be true but most likely reflects the average expectations of a pilot if he cares for the evaluation and tries to meet the qualifications.[125] Hence he might consider a better subjective performance evaluation than the previous year as a friendly action because the superior appreciated his work. If, however, the subjective performance evaluation is worse than the previous year, the pilot might consider it as an

[124] I actually conducted further treatments. As I did not find any differences in the results to these treatments, I used the treatment for which I gained data for all pilots as Player B.

[125] This assumption is further supported by the study of Mas (2006), who showed that police officers final offer served them as a reference point for their expected wage. Also Kahneman, Knetsch and Thaler (1986:730) claimed: „The current wage of an employee serves as reference for evaluating the fairness of future adjustments of that employee's wage." Finally the literature about "psychological contracts" supports the importance of expectations (see e.g. Herriot, Manning and Kidd, 1997).

unfriendly action.[126] If a pilot is reciprocally motivated, he may now reconsider the amount of extra duties he performs. A respective increase or decrease not only helps or hurts the employer but equally important does not affect the pilot's next performance evaluation. Because I already found in chapter three that pilots spend their time on extra-duties as a function of the differences between their last two performance evaluations, I now proceed by a further refinement of those results: Following Kube, Maréchal and Puppe (2006) I further distinguish between positive and negative reciprocity. I thus propose for the same pilots:

- ***Proposition 1a:*** *A pilot increases his time spent on extra duties for one period, if he receives a better subjective performance evaluation than in the previous year.*
- ***Proposition 1b:*** *A pilot decreases his time spent on extra duties for one period, if he receives a worse subjective performance evaluation than in the previous year.*

Having prepared the ground, I now turn to the main subject of this chapter: It seems reasonable to assume that the proposed real life behaviour is actually based on the same kind of individual inclination towards positive and negative reciprocity as I am able to measure in experiments. Hence the measured behaviour in the experimental situation should be able to predict the behaviour in the real life situation. I consequently propose for the positive and the negative case the following:

- ***Proposition 2a:*** *Depending on the experimentally measured positive reciprocity, a pilot increases his time spent on extra duties for one period, after he received a better subjective performance evaluation than before.*
- ***Proposition 2b:*** *Depending on the experimentally measured negative reciprocity, a pilot decreases his time spent on extra duties for one period, after he received a worse subjective performance evaluation than before.*

To test these propositions I make use of my two datasets for the same pilots.

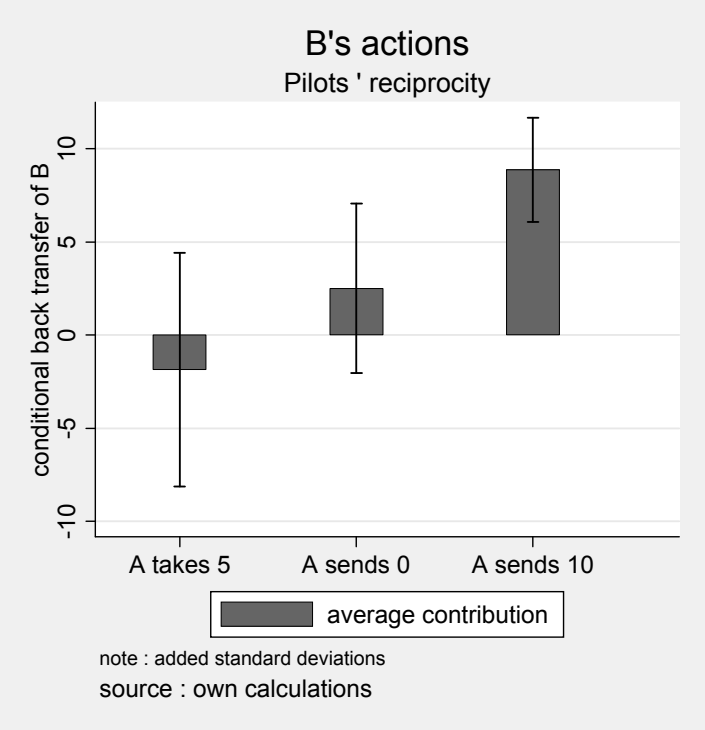
[126] I will of course take into account that the probability of receiving a better evaluation than an already very good evaluation is low. I will discuss its operationalisation in the data section.

4.5. Empirical Analysis

4.5.1 Data

I start this section with the data gained by the experiments in chapter two.[127] I however only discuss the data for the pilots I am able to match with the empirical data of chapter three.[128] A graphical analysis of the measured reactions of the pilots as Player B can be found in Figure 16.

Figure 16: The pilots' reactions to different actions of Player A.



More detailed results are displayed in Table 20 in the appendix. Under the names *posreci* and *negreci* it shows the means and standard deviations of the reactions of the pilots to friendly and unfriendly actions. I consider these reactions as a direct measure for their positive and

[127] A detailed discussion of this data is available in chapter two where I used further experimental results to find out whether pilots differ from students and whether pilots treat outsiders differently to insiders. I found that pilots are significantly more positive reciprocal than students and that pilots do not care who they are paired with.

[128] I was able to match 58 out of 96 pilots. The reduced number of pilots is mainly caused by pilots who participated in the experiment in December 2005 but where not employed yet by the Swiss Air Force in the period from 2001 until 2004.

negative reciprocity. I do not include the results from the neutral setting as they may be biased by the expectations of the pilots in the experiment and do not render me any additional information. It was however important for Player B to know that Player A had the option to be neutral. The results for positive and negative reciprocity show the expected signs. Especially the positive reciprocity seems to be rather high on average, a phenomenon that may be caused by the pre-screening, training and socialization of the pilots for teamwork abilities.[129]

The empirical data used here was provided by the Commander of all professional Swiss Air Force pilots at the end of 2005.[130] Again I am only to discuss the data I am able to match to the experimental data of chapter two.[131] The data consist of individual time recordings of the pilots for two evaluation periods. They start as of November 2002 and end by October 2004. Each pilot had to declare the amount of hours he spent on different tasks in a daily column of a personal monthly excel sheet, and had to submit his file at the end of every month to headquarters, which sometimes used it for the statistical evaluation of the different tasks. This means that the entered data had no consequences for the pilot himself, a fact that should minimize a self-declaration bias.[132] I combine this data with the personal results of the subjective performance evaluations of October 2001, October 2002, October 2003 and October 2004. This procedure renders me the possibility to test my propositions and furthermore allows me to check whether the accomplishment of extra duties does not indeed affect the later performance evaluation. To be able to do this, I had to bring the many excel-files into a format I could analyse. Therefore I first imported the data into statistical software, then generated missing values for impossible entries[133] and finally adapted daily to hourly

[129] Furthermore it seems possible that their daily work environment influences their preferences. Henrich et al. (2001) found that especially people with a high payoff from cooperation in their everyday life are very cooperative. For a further discussion and the comparison of the pilots with students see chapter two.

[130] For a detailed description of the data see also chapter three.

[131] This restriction does not seem critical, as it was strictly coincidental whether a pilot was able to attend the compulsory workshop where the experimental data were gained.

[132] The only problem I had to cope with was incomplete datasets. If a pilot declared in less than 75% of the available standard working time per month the nature of his activities (including days off and sick leave), I treated it as a missing value. To test whether this affected the later results, I run the regressions of Table 13 without this restriction. The results were almost not affected.

[133] I treated daily working hours of less than 0 or more than 24 as impossible. Furthermore I deleted those pilots, who always filled out their excel sheets formally with a daily 8.5 hours.

data by, for example, substituting 8.5 hours off-time for a day off.[134] I then generated an individual yearly average for monthly working hours on each task. Next, I combined all normal pilots' tasks into one single variable called *normal duty* and all extra tasks into another single variable called *extra duty*. The original tasks are listed in Table 21 in the appendix. I finally ended up with one single observation per year per person for *extra duties*. The respective descriptive statistics are displayed in Table 20 in the appendix.

I then combined these observations with the subjective performance evaluation received at the end of the period. In this case I called the variable for the subjective performance evaluation *LOBE,* as this is the internal name for the system. I also combined the topical data with the difference of the last two subjective performance evaluations and called it *LOBElagdelta*. For the period of 2004, for instance, I subtracted the LOBE-evaluation of the period 2002 from the LOBE-evaluation of the period 2003. Specified descriptions of these and all further variables are available in Table 12.

Table 12: Descriptions of the variables sorted by the respective evaluation period.

Period	Variable Name	Description
Evaluation Year 03 (Nov 02 - Okt 03)	Extra Duty	The average of monthly extra hours worked between November 2002 and October 2003
	LOBE	The LOBE-evaluation received by October 2003 (=LOBE03)
	LOBElag	The LOBE-evaluation received by October 2002 (=LOBE02)
	LOBElagdelta	The Difference between the LOBE-evaluation received by October 2002 and October 2001 (=LOBE02-LOBE01)
	posLOBElagdelta	The positive side of LOBElagdelta [=(LOBElagdelta if LOBElagdelta>=0, 0 if LOBElagdelta<0)* its inverse probability of appearance]
	negLOBElagdelta	The negative side of LOBElagdelta [=(LOBElagdelta if LOBElagdelta<=0, 0 if LOBElagdelta>0)* its inverse probability of appearance]
	posreci	The reactions of the pilots to a friendly behaviour of players A in the laboratory experiment
	negreci	The reactions of the pilots to a unfriendly behaviour of players A in the laboratory experiment
	posreci*posLOBElagdelta	The interaction of posreci with posLOBElagdelta
	negreci*negLOBElagdelta	The interaction of negreci with negLOBElagdelta
Evaluation Year 04 (Nov 03 - Okt 04)	Extra Duty	The average of monthly extra hours worked between November 2003 and October 2004
	LOBE	The LOBE-evaluation received by October 2004 (=LOBE04)
	LOBElag	The LOBE-evaluation received by October 2003 (=LOBE03)
	LOBElagdelta	The Difference between the LOBE-evaluation received by October 2003 and October 2002 (=LOBE03-LOBE02)
	posLOBElagdelta	The positive side of LOBElagdelta [=(LOBElagdelta if LOBElagdelta>=0, 0 if LOBElagdelta<0)* its inverse probability of appearance]
	negLOBElagdelta	The negative side of LOBElagdelta [=(LOBElagdelta if LOBElagdelta<=0, 0 if LOBElagdelta>0)* its inverse probability of appearance]
	posreci	The reactions of the pilots to a friendly behaviour of players A in the laboratory experiment
	negreci	The reactions of the pilots to a unfriendly behaviour of players A in the laboratory experiment
	posreci*posLOBElagdelta	The interaction of posreci with posLOBElagdelta
	negreci*negLOBElagdelta	The interaction of negreci with negLOBElagdelta
	Delta Extra Duty	The Difference between the time spent on EXTRA duties in Nov 03-Oct 04 and in Nov 04 - Oct 03.(=Extra04-Extra03)

Notes:
Source: own describtions based on data provided by the Swiss Air Force.

[134] I also used the individual average of daily working hours but found no differences in the following results.

I further generated a variable that only captures positive differences in the last two subjective performance evaluations and called it *posLOBElagdelta*. I did the same for the negative case and generated *negLOBElagdelta*. These last two variables take on the value of zero if the other one is positive, and positive values otherwise. I will use these variables to assess whether a friendly or an unfriendly action has been taken. As it becomes more and more difficult for a superior to issue a better evaluation if the prior evaluation is already good, I applied to the positive and negative differences of the last two evaluations a probability distribution. In this vein I tried to additionally value the friendliness of the action. I estimated the probability of receiving a better evaluation by using the forced distribution applied by the Swiss Government (Swiss Government, 2001:21, Art.49). This means that I multiplied the positive difference of the last two evaluations with the inverse probability of it to be positive at all. This results in a higher value of an increase of a good subjective performance evaluation than of an increase of a bad subjective performance evaluation. Consequently the more unlikely event is of higher value. For negative differences of the last two evaluations I did the opposite, which results in a worse unfriendly action if an already bad subjective performance evaluation is decreased than if a good subjective performance evaluation is decreased.

Combining the experimental and the empirical data, I end up with 116 datasets for 58 pilots.[135]

4.5.2 Procedure and Results

Before I can turn to test the propositions I have to assess whether the results of chapter three are still valid for the restricted dataset. I start by analysing whether the amount of extra duties exerted is actually an unobserved effort level. To verify this assumption I run a standard OLS-regression and a fixed effects regression.[136] The results remain unchanged to the ones in chapter three: According to a fixed effects regression I do not find any positive effect of the

[135] As I am going to use fixed effects regression techniques it is critical that my data bears inter-temporal differences. This should be the case as later shown in Figure 17. Out of the 58 pilots only 12 show zero differences in LOBElagdelta between 2004 and 2003 and there are no pilots that performed in 2003 and 2004 the same amount of extra duties.

[136] I use the evaluation at the end of a year as the dependent variable, while the average amount of extra and normal duties accomplished during the year are the explanatory variables. Additionally I use a dummy-variable for the second evaluation period (*Nov03-Oct04*) to allow for yearly changes in the environment. According to Table 20 in the appendix the LOBE-evaluations do not comprise any corner-solutions, which makes it unnecessary to treat this potential problem.

performed extra duties in the subjective performance evaluation, while in the OLS-regressions the effect is still positive.[137] I hence derive that there are some pilots who perform many extra duties and at the same time receive good evaluations. They, however, do not receive better evaluations if they perform even more extra duties in one year than in another.[138] Indeed, according to the fixed effects estimation I cannot reject the hypothesis that the amount of extra duties accomplished does not influence the later subjective performance evaluation. Hence changes in the quantity of extra duties may be considered as unobserved changes in effort levels.

I now turn to the confirmation of the main result of chapter three, where I found extra duties to depend on the difference of the last two subjective performance evaluations. This means that a pilot performs more extra duties if he is positively surprised and less if he feels disappointed. I thus estimate again with the restricted dataset the effect of the difference of the last two evaluations *(LOBElagdelta)* on the hours a pilot works on average for extra duties per month. I use, for example, the evaluations of October 2002 and October 2003 to assess whether an increase in the evaluation had an effect on the amount of extra duties performed between November 2003 and October 2004. To correct for increased extra duties due to promotions into management positions I include the wage bracket of the pilots as a control variable. Additionally I include a dummy variable for the second period to allow for different external conditions (*Nov03-Oct04*). I start by using the OLS regression, which showed statistically significant results in chapter three at the 10% level.

According to the new results in the first column of Table 13, I observe that the effect of the difference of the last two evaluations on extra duties is positive but not significant anymore. This result may cast doubts on the viability of the data. Two empirical reasons can however be relevant: First, the reduction in the sample size from 192 to 116 observations may have caused the standard error to grow from 5.981 to 9.058.[139] Second, the regression of the now smaller sample could be biased: The OLS-regression only measures whether the pilots who received high increases in their evaluations perform more extra duties than those who did not. But I already know that some pilots receive good evaluations and at the same time perform

[137] Because I am using a smaller sample than in chapter three, it is not surprising that I still do not find any effect in the fixed effects regression.

[138] To further test this issue I rerun the same OLS regression with only those observations that had different LOBE-values for the years 2003 and 2004. The result remains almost the same. Even the significance of the estimator for extra-duty in the OLS regression only climbs by a mere 0.2%.

[139] The point estimate only changed from 10.817 to 9.017.

many extra duties (but not because of the increased evaluation). If in the smaller sample these pilots receive by coincidence less increases in their evaluations than the pilots who do not many extra duties, the result for the OLS regression would show the following: Pilots that did not do many extra duties but received high increases in their evaluations, therefore perform more extra duties than before, still perform less extra duties than the "average" pilot that did not receive a better evaluation.

Table 13: Regression of the effect of positive and negative differences of the last two LOBE-results on the amount of extra duties worked, without and with interacted reciprocity.

Extra duty monthly hours	OLS	Fixed Effects	OLS	Fixed Effects	OLS	Fixed Effects
LOBElagdelta	9.017	16.916				
	(9.058)	(12.529)				
posreci					0.659	
					(0.908)	
posLOBElagdelta			-20.177	75.374 *	-92.931	-175.518
			(31.558)	(42.377)	(118.732)	(107.388)
posreci*posLOBElagdelta					7.722	27.433 **
					(12.601)	(12.267)
negreci					0.784	
					(0.495)	
negLOBElagdelta			-87.843 **	19.036	-95.262 ***	19.366
			(34.823)	(52.461)	(31.157)	(37.740)
negreci*negLOBElagdelta					-7.494 **	-8.105 **
					(3.335)	(3.668)
Wage bracket	11.963 ***	1.097	11.720 ***	-0.118	11.799 ***	-0.501
	(2.413)	(2.045)	(2.224)	(2.644)	(2.114)	(2.922)
Nov03-Oct04	0.782	6.256	-0.243	8.577 *	0.115	10.370 **
	(3.620)	(5.057)	(3.653)	(4.844)	(3.908)	(5.086)
const	-244.515 ***	19.108	-234.501 ***	42.664	-243.312 ***	51.649
	(59.042)	(48.904)	(55.070)	(62.747)	(53.104)	(69.898)
adj. R^2	0.3513	0.7177	0.3772	0.7273	0.4052	0.7498
n	116	116	116	116	116	116
clusters	58	58	58	58	58	58
F-Test for posreci*posLOBElagdelta negreci*negLOBElagdelta						3.97 0.0249

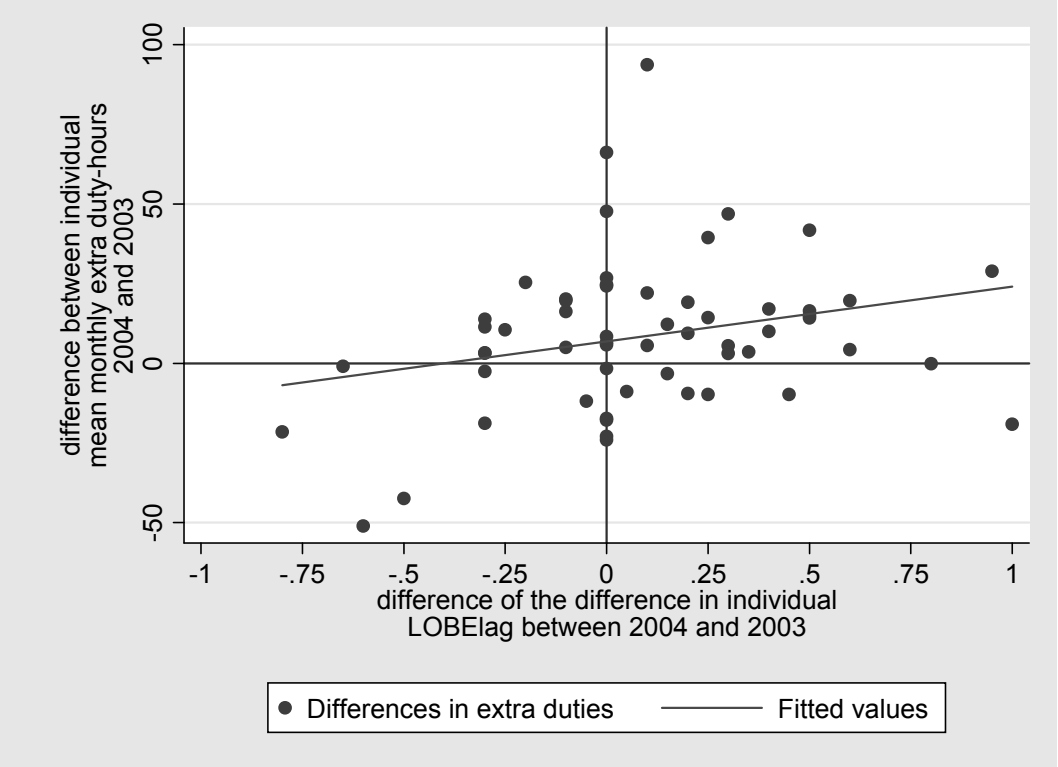
Notes:
Dependent variable: yearly mean of monthly extra hours
Coefficient of Fixed-Effects regressions
(Robust standard errors in OLS and FE-regressions adjusted for clustering on individuals in parentheses).
Level of sign.: ***:p<=0.01,**: 0.01<p<=0.05, *:p<=0.10
Source: own calculations based on Swiss Air Force data.

To rule out this possible bias I proceed by using an individual fixed effects regression and adjust the standard errors for clustering on individuals.[140] Like this I am able to assess, whether changes in the evaluations cause changes in individual behaviour. The results in the second column of Table 13 are this time closer to the ones in chapter three, where I found a

[140] With the use of adjusted standard errors, I prevent them to be to small for the case that the error-term is correlated within the observations of an individual pilot.

point estimate of 14.004 with an adjusted standard error for clustering on individuals of 7.937.[141] This time the point estimate is 16.916 and the adjusted standard error climbs to 12.529, which could be caused again by the shrunken sample size. Nevertheless, the effect of the difference of the last two evaluations on extra duties remains statistically significant according to a one-sided test on the 10% level. A respective graphical analysis is displayed in Figure 17.

Figure 17: Pilots' individual reactions to changes in their subjective performance evaluations.



On the x-axis the fixed-effects nature is captured by the difference in the differences between the last two evaluations. On the y-axis the differences between the amount of performed extra duties in the evaluation periods of 2004 and 2003 are displayed. The fitted values present a positive dependence of the two axes. The figure thus shows that the empirical findings do not seem to be based on coincidence. I prefer to avoid making conclusions about the bias in the OLS-estimation until later, at the conclusion of this section. Yet I infer that my main result of chapter three is to a reasonable extent confirmed.

[141] It is important to note that I report the full r-squared for the case of the fixed effects regression and not only the "within"-value. The respective high value is hence also explained by the explanatory power of the fixed effects themselves.

I thus proceed by empirically testing my first propositions and divide my findings of chapter three into two parts: In an OLS and fixed effects regression I replace the variable for the difference between the last two subjective performance evaluations by it's positive and negative part, namely by positive differences in evaluations *(posLOBElagdelta)* and negative differences in evaluations *(negLOBElagdelta).* The results of the respective estimations are displayed in the third and forth columns of Table 13. It is surprising to see that on the one hand in the OLS regression I find a statistically significant effect of negative differences on the amount of extra duties performed, while the positive differences are not at all statistically significant and even have an unexpected sign; and on the other, in the fixed effects regression, I only find effects for positive differences of the last two evaluations, while in this case negative differences are not statistically significant and bear the wrong sign. That means that the positive result of the OLS regression in the first column of Table 13 was mainly driven by negative differences in evaluations, while the result of the fixed effects regression in the second column of Table 13 was mainly driven by positive differences in evaluations. This could only indicate that the OLS-regression is biased despite the fact it would confirm proposition *1b*.[142] Because already the OLS-coefficient for the effect of extra duties on the later subjective performance evaluations seems to be biased, I focus on the fixed effects regression of column four of Table 13: According to this analysis the positive difference between the last two subjective performance evaluations bears a positive effect on the amount of extra duties performed on a statistical significance level of 10%.

[142] The following situation could induce this bias: As previously seen, the pilots, who on average perform more extra duties, receive also on average better evaluations (but due to a different reason). If the changes in the evaluations appear primarily for lower evaluations, the following situations materializes: According to the fixed effects regression in the forth column of Table 13, increased evaluations lead to increased amounts of performed extra duties while decreased evaluations have no effects. Hence, the pilots, who did not perform many extra duties, but received a better evaluation, now perform a bit more extra duties, but probably still less than the pilots, who started by doing many extra duties but did not receive a changed evaluation. That means that for this example, the pilots who received better evaluations than before perform still less extra duties than the average pilot who did not receive a better evaluation. These coherences might lead to the negative estimator for positive changes of evaluations in the OLS-regression in column three of Table 13. If also only pilots who originally perform few extra duties and receive bad evaluations received negative changes in their evaluation, an equal bias of the OLS-estimation prevails: Even though these pilots may not have changed the amount of extra duties they perform, the pilots who received a worse evaluations now perform on average less extra duties than the average pilot who did received the same evaluation than before.

Hence I cannot reject the hypothesis that a positive difference in the last two subjective performance evaluations augments the average time spent for extra duties. Thus proposition *1a* is supported.

- ***Result 1a:*** *A pilot increases his time spent on extra duties for one period, if he receives a better subjective performance evaluation than the year before.*

For the case that the pilot receives a worse evaluation than the year before, I find however no effects. The estimator in the forth column of Table 13 is not only statistically insignificant but even bears the wrong sign. Hence I must reject the hypothesis that a negative difference in the last two subjective performance evaluations has a negative impact on the time spent for extra duties.

- ***Result 2b:*** *A pilot does not decrease his time spent on extra duties, if he receives a worse subjective performance evaluation than last year.*

This finding is in my view very astonishing. Charness and Rabin (2002) showed that especially negative reciprocity is a strong behavioural force and Mas (2006) found very strong signs of negative reciprocity after police officers had lost final offer arbitration. Also Kube, Maréchal and Puppe (2006:7) report findings of very strong negative reciprocity in their field experiment. I, however, find only positive reciprocity according to my fixed effects regressions. If I would not have had any experimental data, I would probably even call theses findings questionable or attribute them to the fact that the pilots did not actually experience wage cuts but only smaller wage increments as expected. It seems possible, though, that the results may be triggered by something else: May be the very strong positive and weak negative average reciprocity of the screened pilots (also see Figure 16 on page 76) influenced the empirical outcome.[143] To test this idea I turn to propositions *2a* and *2b*.

I proceed by integrating into the regression analysis the results of my experiment. That means that I additionally use two interaction terms: The first term concerns the reaction to friendly actions: In addition to the simple variable for positive differences between the last two subjective performance evaluations, I interact the same variable with the measured positive reciprocity in the experiment (*posreci*posLOBElagdelta*). If proposition *2a* is true, the interaction term should become significant while the sole term for the positive difference of the evaluations (*posLOBElagdelta)* should become statistically insignificant. To test at the

[143] I also showed in chapter two that pilots are far more positive reciprocal than students.

same time proposition *2b* I integrate the interaction of negative reciprocity with the negative differences in the last two subjective performance evaluations (*negreci*negLOBElagdelta*).

To show the full property of my data I start again by an OLS-regression and include the individual values for positive and negative reciprocity. I find the following: According to the fifth column of Table 13, the interaction-terms for the positive and negative reciprocity in the experiment with the real life situation bear the correct signs. However, only the negative term is statistically significant at the 5% level, while I find no significant results for positive reciprocity. Even though these findings are tentative in the light of found negative reciprocity in former studies, it includes one puzzling result: The estimator for the reaction to negative changes in evaluations remains statistically significant and negative like in the regression without the interactions in column three. This casts some doubts on either the validity of my propositions or the applicability of the empirical method.[144] As I most probably experience the same bias in the OLS-regression as before, I proceed again by fixed effects regression techniques.

Hence I turn now to the results of the regression in the last column of Table 13, in which I believe more strongly. In these results, the interaction term for positive reciprocity with the positive differences between the last two subjective performance evaluations is positive and statistically significant at the 5% level. Furthermore the sole positive difference between the last two evaluations is not statistically significant anymore and even becomes negative. Hence I cannot reject that the measured positive reciprocity is able to predict the degree to which a pilot shows positive reciprocity. Furthermore I cannot reject that pilots who do not show any positive reciprocity in the experiment do not show any positive reciprocity in my empirical data neither.

- ***Result 2a:*** *The time a pilot spends on extra duties is a direct function of the interaction of the positive difference of the last two subjective performance evaluations with his measured positive reciprocity.*

[144] I wondered whether it might be the strong effect of the wage brackets that biases my regressions. Even though it seems probable that a squadron commander does according to the estimation in column five 11 hours more extra duties than a normal pilot per month as he has to perform for example the planning of his personnel, this effect could bias my regressions. I therefore performed another regression with fixed effects on wage brackets that however did not provide any new insights. Only seven of my pilots changed wage brackets and all results remain almost the same. Only the estimator for the negative changes in evaluations as well as the estimator for the respective interaction term becomes slightly less significant.

As opposed to my last fixed effects regression, this time also negative reciprocity prevails: According to the last column in Table 13, I find that negative reciprocity is relevant. The interaction term of negative reciprocity with the negative difference in the last two subjective performance assessments is statistically significant at the 5% level. The sole term for negative differences is however not statistically significant as in result *1b*. Hence I cannot reject the thesis that the measured negative reciprocity is able to predict the degree to which a pilot shows negative reciprocity. Furthermore I cannot reject the thesis that pilots who do not show any negative reciprocity in the experiment do not show any negative reciprocity in my empirical data neither.

- ***Result 2b:*** *The time a pilot spends on extra duties is an inverse function of the interaction of the negative difference of the last two subjective performance evaluations with his measured negative reciprocity.*

This result may explain result *1b* as well. As only few pilots are strongly negative reciprocal, they were not able to trigger the result in the forth column of Table 13.

Because of the importance of the results *2a* and *2b* I further used an *F*-test for the two interaction-terms to find out if they are indeed of any importance for the regression. I found an *F*-value of 3.97. I hence cannot reject my combined hypothesis of propositions *2a* and *2b* at a statistical significance level of 2.5%.

My second test concerns the affiliation of the original tasks to normal and extra duties. As it may be argued that the only reason for my findings is a smart sorting of the tasks in Table 21 (see appendix) into normal and extra duties I did the following: I changed the affiliation of the first task by deleting it from his original category and adding it to the other category. Then I rerun the last regression of Table 13.

Thereafter, I reassigned it to its original category and repeated the procedure for the next task. Each time, a significance level of one of the two important estimators drops below 10% in the new regression, I report the specific task in Table 14 together with the new results for the two interaction terms. I find different significance levels for twelve of all tasks if I change their affiliation to the opposite duty. Looking at the descriptions of the tasks reported in Table 14 it seems clear that they have been well assorted. Therefore the results of this chapter appear to be based on correct affiliations of the tasks.

Table 14: The twelve critical tasks for the analysis.

		param.		se	Description
Regression results	posreci x posLOBElag	27.433	**	(12.267)	original regression
(6. column Table 3)	negreci x negLOBElag	-8.105	**	(3.668)	

removed from normal duties added to extra duties		param.		se	Description
auslandkam~v	posreci x posLOBElag	28.118	***	(10.386)	training missions
	negreci x negLOBElag	-6.751		(4.584)	abroad for jets
nightwayni~h	posreci x posLOBElag	26.928	**	(12.186)	specific training mission
	negreci x negLOBElag	-6.404		(3.900)	abroad for jets
nomadflieg~h	posreci x posLOBElag	27.944	**	(12.539)	specific training mission
	negreci x negLOBElag	-6.478		(3.989)	abroad for jets
Flightinstructor	posreci x posLOBElag	24.904		(19.256)	flight instructor
(v46)	negreci x negLOBElag	-8.325	**	(3.866)	
weitereaus~v	posreci x posLOBElag	24.413		(16.544)	other training missions
	negreci x negLOBElag	-8.585	**	(3.655)	abroad for jets
ukfa18	posreci x posLOBElag	21.110		(12.630)	flight training
	negreci x negLOBElag	-8.951	**	(4.126)	on F/A 18
flugdienst~g	posreci x posLOBElag	33.163		(26.619)	individual
	negreci x negLOBElag	-9.764	**	(4.381)	flight training
orgplanung~b	posreci x posLOBElag	25.470		(15.810)	organisation and planing
	negreci x negLOBElag	-8.174	*	(4.256)	of individual flights
flugdienstit	posreci x posLOBElag	26.470	*	(15.018)	individual
	negreci x negLOBElag	-3.109		(6.926)	flight training
spukspugin~n	posreci x posLOBElag	25.576	*	(13.195)	flight instructor
	negreci x negLOBElag	-0.478		(6.076)	for adv. Helic. training
truppendie~d	posreci x posLOBElag	29.523	*	(17.346)	flight training
	negreci x negLOBElag	-4.215		(6.001)	with national reserve

removed from extra duties added to normal duties		param.		se	Description
fhrungorga~n	posreci x posLOBElag	28.762	**	(13.686)	management and
	negreci x negLOBElag	-6.488		(4.505)	organisation

Notes:
Critical Coefficient for Subjective Performance Evaluation (LOBE) of Fixed-Effects regression after changing the category of every single variable.
Dependent variable: yearly mean of monthly extra hours
(Robust standard errors adj. for clustering in parantheses).
Level of sign.: ***:p<=0.01,**: 0.01<p<=0.05, *:p<=0.10
Source: own calculations based on Swiss Air Force data.

4.6. Conclusions

It is still widely discussed whether reciprocal behaviour plays a role in actual labour market situations. Even though many theoretical and experimental studies have addressed this issue, studies using real labour market data are still scarce and their relevance uncertain. To fill this gap I researched within a real firm whether the subjective characteristic of subjective performance evaluations give rise to reciprocal behaviour. Furthermore, I researched whether

I might be able to predict the outcomes with laboratory experiments. Hence I not only sought for a real labour market situation where reciprocity might play a role but also wanted to confirm its validity by experimental data.

I base my research on the two previous chapters: In chapter three I found that the pilots of the Swiss Air Force show reciprocal behaviour as a function of their subjective performance evaluations. I discovered the reciprocal reaction in an effort level that was clearly unobserved by a direct superior, while I had the rare opportunity to measure it. This is of central interest because I strongly assume that in every job such unobserved effort levels exist, though normally nobody is able to measure them. I found that the level of this specific effort provided by a pilot directly depends on the difference between his last two subjective performance evaluations. In chapter two I furthermore researched the individual parameters for reciprocity of the same pilots by experiments and found that they are far more positively reciprocal than their student counterparts.

My main findings are the following: The individual experimental results for positive reciprocity are good predictors for individual reactions to exceeded salary expectations, while individual experimental results for negative reciprocity are good predictors for individual reactions to respective disappointments. This means that the more a worker shows reciprocity in the experiment, the more he shows reciprocity in the labour market situation. I conclude that I indeed found reciprocity in a real labour market.

I furthermore derive that it proved crucial to combine the experimental and the empirical data as each single dataset was not enough to draw a clear picture: The sole empirical data for the effects of "friendly" and "unfriendly" actions only showed that pilots exert more extra duties if their expectations about their subjective performance evaluations are exceeded, while the reactions to disappointments remained undetected. Looking at this preliminary empirical analysis, one could have interpreted that only a positive deviation to the former subjective performance evaluation is relevant.[145]

Even with the respective knowledge of the experimental results of chapter two it would not have been clear whether these findings were able to predict any real behaviour in workers. The pilots could have seen the experiment despite its financial consequences as a simple game

[145] General knowledge gained by experiments with students would have raised questions. Even controlling for social background variables as suggested by Fehr et al. (2003), Gachter, Herrmann and Thoni (2004), Bellmare and Kroeger (2005) and Gueth, Schmidt and Sutter (2007), would not have revealed that Swiss Air Force pilots are that much more positive and almost not at all negative in their reciprocal behaviour.

with no real relevance. Exclusively the one-to-one combination of the two different datasets revealed the complete picture: A pilot truly has to be positively or negatively reciprocal to be affected by a change in his subjective performance evaluation. And also any deterioration of the subjective performance evaluation may be relevant. If a pilot is negatively reciprocal he will retaliate against the worse evaluation by providing a lower unobserved effort level in the next period. Hence the combination of both the experimental with the real life data showed that real life data does not always tell a complete story. With real life data on the one hand one cannot capture more than an average behaviour of a specific group of individuals. With experimental results on the other, one might not be sure whether to be able to predict if a specific real life behaviour exists.

The combination of the data provided me with three insights: Firstly, reciprocity actually matters in real labour market situations. Even if a subjective performance evaluation is applied to forgo the incomplete view of direct incentive systems, employers will never be able to evade dysfunctional behavioural responses of their employees altogether. The negative reciprocal effects of disappointments seem to be of a central issue. Secondly, I might add that it seems more than ever important to screen, train and socialize a workforce intensively for occupations where trustworthiness and reciprocity plays an important role. Hence I emphasize as my third and final point that especially for these jobs, the internal labour market is of utmost importance. If a firm is actually able to screen, socialize and train its workers correctly, it cannot afford to loose them anymore.

CHAPTER 5:

Final Remarks

With this work I wanted to study whether reciprocal preferences found in experiments are important for real internal labour market relations. And indeed, by combining experimental, operational and personnel data, I not only find that reciprocity matters but also that reciprocal real life behaviour is predictable by experiments.

Evidence from my experiments with Swiss Air Force pilots shows that reciprocity may be relevant for the recruiting decisions of employers. I find that pilots are significantly more positively reciprocal and slightly less negatively reciprocal than when compared with students and the general population. Further results even show that pilots treat superiors and even outsiders just as they treat their fellow cockpit-companions. Of course, this behaviour could be based on pure self-selection into this special profession. However, I propose that the Air Force screens, trains and socializes its pilots to enable an optimal collaboration between them because in this environment team-responsibilities are high and supervision is difficult.

I furthermore find that reciprocal preferences also play a role for the relationship between employees and employers. My empirical analysis of real personnel and operational data shows that changes in subjective performance evaluations are reciprocated by higher or lower unobserved effort levels. As the provision of unobserved effort can not be explained by strictly selfish preferences, considerations of reciprocity remain a possible explanation. Interestingly, it is not the absolute values of subjective performance evaluations that seem to matter, but instead, it is the relative values of these evaluations in relation to those of the previous year that matter most. It thus seems as if expectations adapt quickly. I conclude that reciprocity is important for the employer-employee relationship and influences effort levels and performance especially in the short-run.

Finally, I find that the reciprocity measures of my experimental and empirical analysis correlate! The employees that show a stronger reciprocal behaviour in the experiments also show a stronger reciprocal behaviour after changes in their subjective performance evaluations in the real world. To my knowledge this is a unique finding born out of the combination of experimental and empirical data of the same employees. This not only enhances the validity of experiments. It also shows that the empirically found real life behaviour may indeed be explained by reciprocal preferences, which emphasizes that reciprocity is in fact relevant for real workplace relations. Moreover I show that the strong positive reciprocity of the pilots may not only help improve their teamwork abilities but may also improve the employer-employee relationship. With this work I consequently show that reciprocity matters even if on many occasions we are unable to directly observe its effects.

Possible implications of these findings are the following: I believe it is good practice to apply prolonged screening and training for teamwork abilities to employees. Many employers could profit from workers who incorporate strong positive reciprocal preferences. This may not only be applicable to environments where failures to achieve agreements in teams may have detrimental consequences. Also other firms could, with boosted positive reciprocity, reduce the necessary supervision of employees. This may even be of more general relevance.

I believe many firms depend to some extent on the performance of unobserved efforts. As this kind of effort seems to be affected by the reciprocity of employees, it is important to know what triggers reciprocal behaviour. Looking at my empirical results, it seems as if primarily relative changes in working conditions matter, because employees seem to adapt quickly to changed conditions such as better salaries. This implies that wage effective subjective performance evaluation systems do not induce lasting reciprocal behaviour. All the more, if the subjective evaluations have to follow a forced distribution, only few employees may earn above average salaries. Superiors may even be forced to evaluate other employees as below average. If numerous employees were to trust in the accuracy of the evaluations and at first provide higher effort levels to receive better evaluations, many may suffer disappointment.[146] Even if the triggered negative reciprocal behaviour fades out quickly, this experience may

[146] Also Becker (1998: 344f) emphasizes that a standard distribution applied to the results of a subjective performance evaluation may be harmful due to the small number of workers in a group, the non-random selection into this group and the forced below average evaluations for half of the employees. Furthermore Kappel and Uschatz (1992: 38) emphasized that it is important that the expectations of the employees are answered.

undermine the incentive characteristics of the system itself.[147] If ultimately no clear performance measures are available, an employer may save money and avoid frustration if he disconnects salary from such performance evaluations.[148] If not, he will have to allow the average evaluation to rise slowly. After some years, however, many of his employees will reach the top evaluation. If he can not promote them into new positions, where the evaluation is adapted to stronger criteria, he will have to replace the whole evaluation system by a slightly different one every few years.[149] Only by this measure he can start the motivating rise of evaluations again.

Further research could investigate carefully the overall effectiveness and cost efficiency of *subjective* performance evaluations and bonus systems. As well, it could try to separate the influence of self selection, screening, training and socialization on the reciprocal preferences of employees. Experiments may be used for predictions. But special care has to be taken to most closely represent the special subgroup of the population and the specific work environment of a specific firm. Better yet, future research should use again the combination of experimental and empirical data.

[147] Brudney and Condrey (1993:141) conducted a survey of a similar qualitative performance system and found that significant factors among others for the system to be effective are its accuracy and fairness.

[148] If instead he introduces a performance evaluation according to measurable tasks, the new evaluation may give raise to inefficiently focused efforts of employees.

[149] According to anecdotic evidence this seems to be a common procedure.

References

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner. 2000. “The Moonlighting Game. An experimental Study on reciprocity and retribution.” Journal of Economic Behavior & Organization, Vol. 42, No. 2 (Jun), pp. 265–277.

Air Transport. 2007. „Kurzprotokoll Informationsrapport Sparte LT 01/07.“ Internal Swiss Air Force Document. Dübendorf.

Aircraft Accident Investigation Bureau (AAIB), 2006. “Report No: u1882.” Bern.

Airforcepilot, 2004. “We expect.” http://www.vbs-ddps.ch/internet/luftwaffe/en/home/airforcepilot/chapter0/chapter0.html from 19.8.2004.

Akerlof, George A, and Janet L. Yellen. 1990. “The Fair Wage-Effort Hypothesis and Unemployment.” The Quarterly Journal of Economics, Vol. 105, No. 2, pp. 255-283.

Akerlof, George A, and Rachel E. Kranton. 2005. „Identity and the Economics of Organizations.” Journal of Economic Perspectives, Vol. 19, No. 1 (Winter), pp. 9–32.

Akerlof, George A. 1982. “Labor Contracts as Partial Gift Exchange.” The Quarterly Journal of Economics, Vol. 97, No. 4 (Nov), pp. 543-569.

Akerlof, George A., and Janet L. Yellen. 1986. “Efficiency Wage Models of the Labor Market.” Cambridge, MA: Cambridge University Press.

Akerlof, George A., and Janet L. Yellen. 1988. “Fairness and Unemployment.” The American Economic Review, Vol. 78, No. 2 (May), pp. 44-49.

Alesina, Alberto and Eliana La Ferrara. 2000. “Who Trust Others.” Journal of Public Economics, Vol. 85, No. 2, pp. 207-234.

Andreoni, James, William Harbaugh, and Lise Vesterlund. 2003. “The Carrot or the Stick: Rewards, Punishments, and Cooperation.” The American Economic Review, Vol. 93, No. 3 (Jun), pp. 893-902.

Backes-Gellner, Uschi, Edward P. Lazear, and Birgitta Wolff. 2001. Personalökonomik, Fortgeschrittene Anwendungen für das Management. Stuttgart: Schäffer-Poeschel Verlag.

Baker, George. 2000. “The Use of Performance Measures in Incentive Contracting.” The American Economic Review, Vol. 90, No. 2 (May), pp. 415-420.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. “Social Preferences and the Response to Incentives: Evidence From Personnel Data.” Quarterly Journal of Economics, Vol. 120, No. 3 (Aug), pp. 917-962.

Bartel, Ann, Richard Freeman, Cassey Ichniowski, and Morris Kleiner. 2004. “Can a Work Organization Have an Attitude Problem? The Impact of Workplaces on Employee Attitudes and Economic Outcomes.” CEP Discussion Paper No. 636. http://cep.lse.ac.uk/.

Becker, Fred G. 1998. “Grundlagen betrieblicher Leistungsbeurteilungen: Leistungsverständnis und –prinzip, Beurteilungsproblematik und Verfahrensprobleme.“ Stuttgart: Schaeffer-Poeschel Verlag.

Becker, Manfred. 1999. “Personalentwicklung: Bildung, Förderung und Organisationsentwicklung in Theorie und Praxis”. Stuttgart: Schaeffer-Poeschel Verlag.

Bellmare, Charles, and Bruce Shearer. 2007. “Gift Exchange within a Firm: Evidence from a Field Experiment.” IZA Discussion Paper No 2696. www.iza.org.

Bellmare, Charles, and Sabine Kroeger. 2005. „On representative social capital.” IZA Discussion Paper No. 1145, www.iza.org..

Bellmare, Charles, Sabine Kroeger and Arthur van Soest. 2005. “Actions and Beliefs: Estimating Distribution-Based Preferences using a large scale Experiment with Probability Questions on Expectations.” IZA Discussion Paper No 1666. www.iza.org.

Benjamin, Daniel J. 2006. „Fairness and Efficiency: The Rotten Firm Theorem.” Unpublished paper, Dartmouth College.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. “Trust, Reciprocity, and Social History.” Games and Economic Behavior, Vol. 10, pp. 122-142.

Bernhard, Helen, Ernst Fehr and Urs Fischbacher. 2006. „Group Affiliation and Altruistic Norm Enforcement.” American Economic Review, Vol. 96, No. 2, pp. 217-221.

Bettinger, Eric, and Robert Slonim. 2005. “Using experimental Economics to measure the effects of a natural educational experiment on altruism.” NBER Working Paper No. 11725. www.nber.org.

Bloom, Matt, and George T. Milkovich. 1998. “Relationships among Risk, Incentive Pay, and Organizational Performance.” The Academy of Management Journal, Vol. 41, No. 3 (Jun), pp. 283-297.

Bolton, Gary E., and Axel Ockenfels. 2000. “A Theory of Equity, Reciprocity and Competition.” The American Economic Review, Vol. 90, No. 1 (Mar), pp. 166-193.

Bonjour, Dorothe. 1997. „Lohndiskriminierung in der Schweiz.“ Bern: Paul Haupt Verlag.

Borjas, George J. 2002. Labor Economics. New York: McGraw-Hill.

Brown, Martin, Armin Falk, and Ernst Fehr. 2002. „Contractual Incompleteness and the Nature of Market Interaction.” Unpublished paper, University of Zurich.

Brown, Martin, Armin Falk, and Ernst Fehr. 2004. „Relational Contracts and the Nature of Market Interactions.” Econometria, Vol. 72, No. 3, pp. 747-780.

Brudney, Jeffrey L., and Stephen E. Condrey. 1993. “Pay for Performance: Explaining the Differences in Managerial Motivation.” Public Productivity & Management Review, Vol. 17, No. 2 (Winter), pp. 129-144.

Burks, Stephen V., Jeffrey P. Carpenter and Eric Verhoogen. 2003. “Playing both roles in the trust game.” Journal of Economic Behaviour & Organization, Vol. 51, pp. 195-216.

Burks, Stephen V., Jeffrey P. Carpenter and Lorenz Goette. 2006. “Performance Pay and the Erosion of Worker Cooperation: Field experimental evidence.” IZA Working Discussion Paper No 2020. www.iza.org.

Burks, Stephen V., Jeffrey P. Carpenter, Lorenz Goette, Kristen Monaco, Aldo Rustichini, Kay Porter. 2007. “Using Behavioral Economic Field Experiments at a Large Motor Carrier: The Context and Design of the Truckers and Turnover Project”. NBER Working Paper No. 12976. www.nber.org.

Camerer, C., Fehr, E. 2002. Measuring Social Norms and Preferences using Experimental Games: A Guide for Social Scientists. Working Paper 97, IEW-Working-Papers…

Camerer, Colin F., and Ernst Fehr. 2006. „When Does "Economic Man" dominate Social Behaviour.” Science, Vol. 311, No. 6 (Jan), pp. 47-52.

Camerer, Colin F., and Robin M. Hogarth. 1999. “The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework.” Journal of Risk and Uncertainty, Vol. 19, pp. 7-42.

Cappelli, Peter, and Keith Chauvin. 1991. “An Interplant Test of the Efficiency Wage Hypothesis.” The Quarterly Journal of Economics, Vol. 106, No 3(Aug), pp. 769-787.

Carpenter, Jeffrey P., Stephen Burks, and Eric Verhoogen. 2005. “Comparing Students to Workers: The effects of social framing on behaviour in distribution games.” In Carpenter, Jeffrey P., GlennW. Harrison, and John A. List, ed. “Field Experiments in Economics, Research in Experimental Economics.” Vol. 10, Greenwich and London: JAI Press, pp. 261-290.

Carter, John R., and Marco Castillo. 2002. “The Economic Impacts of Altruism, Trust and Reciprocity: An Experimental Approach to Social Capital.” Unpublished paper, University of Wisconsin.

Carter, John R., and Michael D. Irons. 1991. “Are Economics Different, and if so, why?” The Journal of Economic Perspectives, Vol. 5, No. 2 (Spring), pp. 171-177.

Cavalluzzo, Linda C. 1991. “Nonpecuniary Rewards in the Workplace: Demand Estimates Using Quasi-Market Data.” The Review of Economics and Statistics, Vol. 73, No. 3(Aug), pp. 508-512.

Charness, Gary, 2004. “Attribution and Reciprocity in an Experimental Labor Market.” Journal of Labor Economics, 22 (3), pp. 665-688.

Charness, Gary, and Matthew Rabin. 2002. “Understanding Social Preferences with Simple Tests.” Quarterly Journal of Economics, Vol. 117, No. 3(Aug), pp. 817-870.

Charness, Gary, and Matthew Rabin. 2002. “Understanding Social Preferences with Simple Tests.” Quarterly Journal of Economics, Vol. 117, No. 3 (Aug), pp. 817-870.

Coupé, Tom, Valérie Smeets, and Frederic Warzynski. 2003. “Incentives in Economic Departments: Testing Tournaments? SSRN Paper No. 499022, http://papers.ssrn.com.

Cox, James C., Klarita Sadiraj, and Vjollca Sadiraj. 2002. “Trust, Fear, Reciprocity and Altruism.” Discussion paper, University of Arizona.

Dohmen, Thomas, and Armin Falk. 2006. “Performance Pay and Multi-dimensional Sorting: Productivity, Preferences and Gender.” IZA Discussion Paper No. 2001, www.iza.org.

Dohmen, Thomas, Armin Falk, David Huffman and Uwe Sunde. 2006. “Homo Reciprocans: Survey Evidence on Prevalence, Behavior and Success.” IZA Discussion Paper No. 2205, www.iza.org.

Egas, Martijn, and Arno Riedl. 2005. “The Economic of Altruistic Punishment and the Demise of Cooperation.” IZA Discussion Paper No 1646, www.iza.org.

Eisenberg, Elliot F. and Patricia W. Ingraham. 1993. „Analyzing the Comparative Pay-for-Performance Experience: Are There Common Lessons?” Public Productivity & Management Review, Vol. 17, No. 2 (Winter), pp. 117-128.

Eriksson, Tor, and Marie-Claire Villeval. 2004. „Other-Regarding Preferences and Performance Pay – An Experiment on Incentives and Sorting.” IZA Discussion Paper No. 1991, www.iza.org.

Eriksson, Tor, Sabrina Teyssier, and Marie-Claire Villeval. 2006. „Self-Selection and the Efficiency of Tournaments.” IZA Discussion Paper No. 1983, www.iza.org.

Fairris, David, and Lee J. Alston. 1994. “Wages and the Intensity of Labor Effort: Efficiency Wages versus Compensating Payments.” Southern Economic Journal, Vol. 61, No. 1 (Jul), pp. 149-160.

Falk, Armin, and Christian Zehnder. 2006. “Trust and Discrimination. A City-Wide Experiment.” Manuscript. University of Zurich.

Falk, Armin, and Simon Gachter. 2002. „Reputation and Reciprocity: Consequences for the Labour Relation.” Scandinavian Journal of Economics, Vol. 104, No. 1, pp. 1-26.

Falk, Armin, and Urs Fischbacher. 2000. „A Theory of Reciprocity.” Unpublished paper. University of Zurich.

Fallgatter, Michael J. 1999. „Leistungsbeurteilungstheorie und –praxis: Zur „Rationalität“ der Ignorierung theoretischer Empfehlungen“. Zeitschrift für Personalforschung, Year 13, No. 1, pp. 82-100.

Falter, Jean-Marc, and Giovanni Ferro Luzzi. 2000. “Public-Privat Sector Wage Differentials in Switzerland.” Schweizerische Zeitschrift für Volkswirtschaft und Statistik, Vol. 136, No. 3, pp. 319-340.

Federal Department of Defence. 2003a. “Verordnung des VBS über das militärische Personal.“ http://www.admin.ch/ch/d/sr/c172_220_111_310_2.html.

Federal Department of Defence. 2003b. “Verordnung des VBS über die Zulagen im Flug- und Fallschirmsprungdienst des VBS.“ http://www.admin.ch/ch/d/as/2003/1271.pdf.

Fehr, Ernst, and Armin Falk. 1999. „Wage Rigidity in a Competitive Incomplete Contract Market.” The Journal of Political Economy, Vol. 107, No. 1 (Feb), pp. 106-134.

Fehr, Ernst, and John A. List. 2004. “The hidden costs and returns of incentives: Trust and trustworthiness among CEO's.” Journal of European Economic Association, Vol. 2, No. 5 (Sep), pp. 743-771.

Fehr, Ernst, and Klaus M. Schmidt. 1999. „A theory of Fairness, Competition and Cooperation.” The Quarterly Journal of Economics, Vol. 114, No. 3 (Aug), pp. 817-868.

Fehr, Ernst, and Klaus M. Schmidt. 2004. „Fairness and Incentives in a Multi-Task Principal-Agent Model.” Scandinavian Journal of Economics, Vol. 106, No. 3 (Oct), pp. 453-474.

Fehr, Ernst, and Lorenz Goette. 2005. „Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment.” Unpublished paper, University of Zurich.

Fehr, Ernst, and Simon Gachter. 2000a. „Cooperation and Punishment in Public Goods Experiments.” The American Economic Review, Vol. 90, No.4, pp. 980-994.

Fehr, Ernst, and Simon Gachter. 2000b. „Fairness and Retaliation: The Economics of Reciprocity.” The Journal of Economic Perspectives, Vol. 14, No. 3 (Summer), pp. 159-181.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl. 1993. „Does Fairness Prevent Market Clearing? An Experimental Investigation.” The Quarterly Journal of Economics, Vol. 108, No. 2 (May), pp. 437-459.

Fehr, Ernst, Simon Gachter, and Georg Kirchsteiger. 1997. „Reciprocity as a Contract Enforcement Device: Experimental Evidence.” Econometrica, Vol. 65, No. 4 (Jul), pp. 833-860.

Fehr, Ernst, Urs Fischbacher, Bernhard von Rosenbladt, Jürgen Schupp, and Gert G. Wagner. 2003. "A Nation-Wide Laboratory: Examining Trust and Trustworthiness by Integrating Behavioral Experiments into Representative Survey" IZA Discussion Paper No. 715. www.iza.org.

Fehr, Ernst., Naef, Michael, Urs Fischacher, Juergen Schupp, and Gert. G.. Wagner. 2006. "Why Americans are more trusting than Germans." Working Paper, University of Zurich.

Fershtman, Chaim, and Uri Gneezy. 2001. "Discrimination in a segmented society: an experimental Approach." The Quarterly Journal of Economics, Vol. 116, No. 1, pp. 351-377.

Fox, Jeremy B., K. Dow Scott, Joan M. Donohue. 1993. "An Investigation Into Pay Valence and Performance in a Pay-for-Performance Field Setting." Journal of Organizational Behavior, Vol. 14, No. 7 (Dez), pp. 687-693.

Frank, Robert H., Thomas Gilovich and Dennis T. Regan. 1993. "Does Studying Economics Inhibit Cooperation?" The Journal of Economic Perspectives, Vol. 7, No. 2 (Spring), pp. 159-171.

Gachter, Simon, Benedikt Herrmann, and Christian Thoni. 2004. "Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence." Journal of Economic Behavior & Organization, Vol. 55, No. 4 (Dez), pp. 505-531.

Gachter, Simon, Esther Kessler, and Manfred Koenigstein. 2006. „Performance Incentives and the Dynamics of Voluntary Cooperation." Unpublished paper, University of Nottingham.

Gachter, Simon, and Ernst Fehr. 1999. „Collective Action as a Social Exchange." Journal of Economic Behavior & Organisation Vol. 39, 341–369.

Gibbons, Robert. 1998. "Incentives in Organizations." Journal of Economic Perspectives, Vol. 12, No. 4 (Fall), pp.115-132.

Gilbert, Daniel T., Elizabeth C. Pinel, Timothy D. Wilson, and Stephen J. Blumberg. 1998. "Immune Neglect: A Source of Durability Bias in Affective Forecasting." Journal of Personality and Social Psychology, Vol. 75, No. 3, pp. 617-638.

Glaeser, Edward L., David Laibson, and Bruce Sacerdote. 2002. "An Economic Approach to Social Capital." The Economic Journal, Vol. 112, No. 483 (Nov), pp. 437-458.

Glaeser, Edward, David Laibson, Jose A. Scheinkman, and Christine L. Soutter. 2000. "Measuring Trust." The Quarterly Journal of Economics, Vol. 115, No. 3 (Aug), pp. 811-846.

Gneezy, Uri, John A. List. 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets using Field Experiments." NBER Working Paper No. 12063, www.nber.org.

Goette, Lorenz, David Huffman and Stephan Meier. 2006. „The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence using Random Assignment to Real Social Groups." American Economic Review, Vol. 96, No. 2, pp. 212-216.

Gueth, Werner, Carsten Schmidt, and Matthias Sutter. 2007. "Bargaining Outside the Lab: A Newspaper Experiment of a Three-Person Ultimatum Game." Economic Journal, Vol. 117, No. 518 (Mar), pp. 449-469.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." Journal of Economic Literature, Vol. 42, No. 4 (Dez), pp. 1009-1055.

Head of the Armed Forces. 2004. „Weisungen zur Handhabung der Lohnentwicklung, der Anerkennungs- und Einsatzprämien im Departementsbereich Verteidigung." Armed Forces Directive.

Hedinger, Beat. 2004. "Employability of Specialists in situations of structural change. A theoretical and empirical Analysis." Diplomathesis, University of Zurich.

Helmreich, Robert L., Ashleigh C. Merritt and John A. Wilhelm. 1999. „The evolution of Crew Resource Management training in commercial aviation." International Journal of Aviation Psychology, Vol. 9, No. 1, pp. 19-32.

Hennig-Schmidt, Heike, Bettina Rockenbach, Abdolkarim Sadrieh. 2005. „In Search of Workers' Real Effort Reciprocity – A Field and a Laboratory Experiment." Unpublished paper, University of Bonn.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." The American Economic Review, Vol. 91, No. 2 (May), pp. 73-78.

Herriot, Peter, W.E.G. Manning, and Jennifer M. Kidd. 1997. "The Content of the Psychological Contract." British Journal of Management, Vol. 8, pp. 151-162.

Hoff, Karla, and Priyanka Pandey. 2006. „Discrimination, Social Identity, and Durable Inequalities." American Economic Review, Vol. 96, No. 2, pp. 206-211.

Hoffmann, Elizabeth., Kevin McCabe, and Vernon L. Smith. 1996. „Social Distance and Other-Regarding Behavior in Dictator Games." The American Economic Review, Vol. 86, No. 3 (Jun), pp. 653-660.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." Scandinavian Journal of Statistic, Vol. 6, pp. 65-70.

Kahn, Lawrence M. and Peter D. Sherer. 1990. "Contingent Pay and Managerial Performance." Industrial and Labor Relations Review, Vol. 43, No. 3 (Feb), pp. 107S-120S.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. „Fairness as a Constraint on Profit Seeking: Entitlements in the Market." The American Economic Review, Vol. 76, No. 4 (Sep), pp. 728-743.

Kahneman, Daniel. 2003. "Maps of Bounded Rationality: Psychology for Behavioral Economics." The American Economic Review, Vol. 93, No. 5, pp. 1449-1475.

Kandel, Eugene, and Edward P. Lazear. 1992. „Peer Pressure and Partnerships." The Journal of Political Economy, Vol. 100, No. 4 (Aug), pp. 801-817.

Kappel, Heinz, Philippe Uschatz. 1992. „Variable Kaderentlöhnung: Erfolg und Leistung honorieren." Zürich: Verl. Industrielle Organisation.

Karlan, Dean S. 2004. „Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. " Unpublished paper, Princeton University.

Katz, Harry C., Thomas A. Kochan, and Kenneth R. Gobeille. 1983. "Industrial Relations Performance, Economic Performance, and QWL Programs: An Interplant Analysis." Industrial and Labor Relations Review, Vol. 37, No. 1 (Oct), pp. 3-17.

Katz, Lawrence F. 1986. "Efficiency Wage Theories: A Partial Evaluation." NBER Macroeconmics Annual, Vol. 1, pp. 235-276.

Kleiner, Morris M., Jonathan S. Leonard, and Adam M. Pilarski. 2002. "How Industrial Relations Affects Plant Performance: The Case of Commercial Aircraft Manufacturing." Industrial and Labor Relations Review, Vol. 55, No. 2 (Jan), pp. 195-218.

Krakel, Matthias. 2004. "Emotions and Incentives." IZA Discussion Paper No. 1270, www.iza.org.

Krueger, Alan B. and Alexandre Mas. 2004. "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires." Journal of Political Economy Vol. 112, No. 2 (Apr), pp. 253-289.

Kube Sebastian, Michel A. Maréchal, and Clemens Puppe. 2006. “Putting Reciprocity to work – Positive versus Negative Responses in the Field”. Unpublished paper, University of St. Gallen.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert W. Vishny. 1997. “Trust in Large Organizations.” The American Economic Review, Vol. 87, No. 2, pp. 333-338.

Lavy, Victor. 2004. “Performance Pay and Teachers’ Effort, Productivity and Grading Ethics.” NBER Working Paper No. 10622. www.nber.org.

Lazear, Edward P. 2000. “Performance Pay and Productivity.” The American Economic Review, Vol. 90, No. 5 (Dez), pp. 1346-1361.

Lazear, Edward P., Ulrike Malmendier, Roberto A. Weber. 2005. “Sorting in Experiments with Application to Social Preferences.” Unpublished paper, Stanford Graduate School of Business.

Levine, David K. 1998. “Modeling Altruism and Spitefulness in Experiments.” Review of Economic Dynamics, Vol. 1, pp. 593-622.

Levitt Steven D., and John A. List, 2007. “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?” Journal of Economic Perspectives, Vol. 21, No. 2 (Spring), pp. 153-174.

List, John A., 2007. “Field Experiments: A Bridge Between Lab and Naturally-Occurring Data”. NBER Working Paper No. 12992. www.nber.org.

Loewenstein, George. 2005. “Hot-cold Empathy Gaps and Medical Decision Making,” Health Psychology, Vol. 24, No. 4 (Suppl.), pp. S49-S56.

Manski, Charles F. 2002. “Identification of decision rules in experiments on simple games of proposal and response.” European Economic Review, Vol. 46, pp. 880-891.

Mas, Alexandre. 2004. “Labor Unrest and the Quality of Production: Evidence from the Construction Equipment Resale Market.” Unpublished paper, University of California Berkeley.

Mas, Alexandre. 2006. “Pay, Reference Points, and Police Performance.” NBER Working Paper No. 12202. www.nber.org.

Noser, Philip. 2003. „Untersuchung des Auswahlsystems für die Fliegerische Vorschulung des Bundes (FVS) aus dem Jahre 2000.“ Diplomathesis, University of Zurich.

Prendergast, Canice. 1999. „The Provision of Incentives in Firms.” Journal of Economic Literature, Vol. 37, No. 1 (Mar), pp. 7-63.

Rajagopalan, Nandini. 1997. “Strategic Orientations, Incentive Plan Adoptions, and Firm Performance: Evidence form Electric Utility Firms.” Strategic Management Journal, Vol. 18, No. 10 (Nov), pp. 761-785.

Romano, Joseph P., and Michael Wolf. 2005. “Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing.” Journal of the American Statistical Association, Vol. 100, No. 469 (Mar), pp. 94-108.

Ruffle, Bradley J. and Richard Sosis. 2006. “Cooperation and the In-Group-Out-Group Bias: A Field Test on Israeli Kibbutz Members and City Residents.” Journal of Economic Behavior & Organization, Vol. 60, No. 2, pp. 147-163.

Schettgen, Peter. 1996. „Arbeit Leistung Lohn. Analyse- und Bewertungsmethoden aus sozioökonomischer Perspektive.“ Stuttgart: Ferdinand Enke Verlag.

Schindler, Roxane. 1997. „Die allgemeine Dienstpflicht.“ Diss., University of Zurich.

Schopenhauer, Arthur. 1851. „Aphorismen zur Lebensweisheit.“ In Rudolf Marx, ed., 1974 Stuttgart: Kröner, pp. 47f, 48f, 8-12.

Schweizerische Unfallversicherungsanstalt Suva. 2004. „Unfallstatistik UVG 1998-2002.“ SUVA, Luzern.

Shapiro, Carl, and Joseph E. Stiglitz. 1984. “Equilibrium Unemployment as a Worker Discipline Device.” The American Economic Review Vol. 74, No. 3 (Jun), pp. 433-444.

Slonim, R., Roth, A., 1998. Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic. Econometrica 66(3), 569-596

Smith, Adam. 1776. „An Inquiry into the Nature and Causes of the Wealth of Nations“, ed. Edwin Cannan, 1904, Methuen and Co., Ltd, London, www.econlib.org.

Sobel, Joel. 2005. „Interdependent Preferences and Reciprocity.” Journal of Economic Literature, Vol. 43 (Jun), pp. 392-436.

Swiss Government. 1995. „Bundesgesetz vom 3. Februar 1995 über die Armee und die Militärverwaltung (Militärgesetz, MG).“ http://www.admin.ch/ch/d/sr/5/510.10.de.pdf.

Swiss Government. 2001. „Bundespersonalverordnung (BPV).“ http://www.admin.ch/ch/d/sr/1/172.220.111.3.de.pdf.

Verhoogen, Eric A., Stephen V. Burks, and Jeffrey P. Carpenter. 2006. “Fairness and Freight-Handlers: Local Labor Market Conditions and Wage-Fairness Perceptions in a Trucking Firm.” IZA Discussion Paper No. 1352. www.iza.org.

Wooldridge, Jeffrey M. 2003. “Introductory Econometrics: A Modern Approach.” Mason, Ohio: Thomson/South-Western Publishing.

Appendix

Table 15: B's absolute spendings conditional different actions of A (Comparison of S-S and P-P treatments)

	B's absolute spending cond. negative action of A	B's absolute spending cond. neutral action of A	B's absolute spending cond. positive action of A
pilot	-0.049	0.084	4.051$^{###}$
	(1.561)	(1.500)	(1.053)***
male	2.560	0.674	2.550
	(1.373)	(1.166)	(1.294)*
age	-0.366	0.182	-0.596
	(0.357)	(0.345)	(0.261)**
age^2	0.005	-0.003	0.007
	(0.005)	(0.005)	(0.003)**
educ	-.062	0.213	0.168
	(0.381)	(0.387)	(0.150)
const.	9.654	-3.519	12.561
	(7.407)	(7.371)	(4.303)***
R^2	0.059	0.020	0.320
Prob > F	0.307	0.907	0.0000
n	88	88	88

***Notes*: Dependent variable: absolute amount spent by B after different actions of A. Coefficients of OLS-regression (Robust standard errors in parentheses).**
***Level of significance:* *0.1<p, **0.01<p<0.05, ***p<0.01 corrected levels for multiple hypothesis (Holm, 1979): $^{#}$0.1<p, $^{##}$0.01<p<0.05, $^{###}$p<0.01**
***Source:* own calculations based on experimental evidence from December 2005.**

Table 16: Actions and beliefs in treatment S-S

Student-Student	**A**	**B**
participants	34	34
missing actions	-	-
antireciprocical actions	-	1
used actions	**34**	**33**
missing beliefs	2	1
beliefs adding up to more than 100%	-	4
used beliefs	**32**	**28**

Source: own calculations based on experimental evidence from 12/05.

Table 17: Actions and beliefs in treatment S-P

Student-Pilot	**A**	**B**
participants	58	116
missing actions	-	-
antireciprocical actions	-	4
used actions	**58**	**112**
missing beliefs	1	3
beliefs adding up to more than 100%	2	13
used beliefs	**55**	**96**

Source: own calculations based on experimental evidence from 12/05.

Table 18: Actions and beliefs in treatment H-L

H-L	**A**	**B**
participants	49	67
missing actions	-	-
antireciprocical actions	-	1
used actions	**49**	**66**
missing beliefs	-	-
beliefs adding up to more than 100%	2	7
used beliefs	**47**	**59**

Source: own calculations based on experimental evidence from 12/05.

Table 19: Actions and beliefs in treatment P-P

Pilot-Pilot	**A(H+L)**	**B(H+L)**
participants	25+34	24+33
missing actions	1+2	1+0
antireciprocical actions	-	0+1
used actions	**24+32**	**23+32**
missing beliefs	-	-
beliefs adding up to more than 100%	1+1	3+3
used beliefs	**23+31**	**20+29**

Source: own calculations based on experimental evidence from 12/05.

Table 20: Descriptive Statistics.

Variable	Year	Obs	Mean	Std. Dev.	Min	Max	Description
Extra Duty	03	58	43.971	30.433	0.000	113.833	The average of monthly
	04	58	52.368	31.679	6.000	127.700	extra hours worked over one qualifciation period
Delta Extra Duty	04	58	8.397	23.911	-51.083	93.677	Difference between Extra Duty 04 and 03
LOBE value	01	58	3.434	0.392	2.700	4.200	LOBE-Evaluation
	02	58	3.336	0.279	2.900	4.100	at the end of a qualification
	03	58	3.328	0.257	3.000	4.000	period
	04	58	3.356	0.267	2.600	4.000	
LOBElag value	02	58	3.434	0.392	2.700	4.200	LOBE-Evaluation
	03	58	3.336	0.279	2.900	4.100	of the last period (t-1)
	04	58	3.328	0.257	3.000	4.000	(LOBElag02=LOBE01)
LOBElagdelta	03	58	-0.098	0.285	-0.900	0.400	Difference of last two LOBE
	04	58	-0.009	0.183	-0.400	0.400	Eval. (LOBElagdelta03 = LOBE02-LOBE01)
posLOBElagdelta	03	58	0.043	0.090	0.000	0.307	The positive side of
	04	58	0.048	0.083	0.000	0.360	LOBElagdelta weighted by its inverse prob. of appearance
negLOBElagdelta	03	58	0.047	0.080	0.000	0.385	The negative side of
	04	58	0.024	0.046	0.000	0.200	LOBElagdelta weighted by its inverse prob. of appearance
posreci	03	58	8.879	2.798	0	10	The experimentaly measured
	04	58	8.879	2.798	0	10	positive reciprocity
negreci	03	58	1.862	6.267	-10	10	The experimentaly measured
	04	58	1.862	6.267	-10	10	negative reciprocity
posreci*posLOBElagdelta	03	58	0.407	0.861	0.000	3.072	Interaction term of
	04	58	0.398	0.761	0.000	3.600	posreci with posLOBElagdelta
negreci*negLOBElagdelta	03	58	0.010	0.734	-3.846	2.040	Interaction term of
	04	58	0.034	0.295	-1.071	1.166	negreci with negLOBElagdelta
Wage bracket	03	58	24.190	1.721	17	28	The wage bracket a worker
	04	58	24.759	1.261	24	29	belongs to

Notes:
Source: own calculations based on data provided by the Swiss Air Force.

Table 21: List of extra and normal duties.

Normalduties

Variablename	Variablename
administration	nightwayinstruktion
aquila	nightwaynichtfliegerisch
ausbildnerfluglehrer	nomadeausbildung
auslandkampagnenressortluv	nomadefliegerisch
auslandkampagnenueg	nomadeinstruktion
avia	nomadenichtfliegerisch
axalp	nomadfliegerisch
bealwausstellunginbern	nomadnichtfliegerisch
besucheausstellungenallg	norkaausbildung
bmpsausbildung	norkafliegerisch
bmpsinstruktion	norkainstruktion
colibri	norkanichtfliegerisch
comptoirsuisse2003	norkanomade
condorusterdtruppe	orgplanungflugbetrieb
crm	patrouillesuisse
expo02werftlw	pc7team
feuerlscheinstze	pikettreadiness
flgebetreuungmiliz	simulatoreigenestraining
flgezgdritterhelipt	simulatorinstruktion
flgezgdritterlainkldo	spesen
flgezgdritterltjetinklska	sport
flgezgdritterltjetinklskaf	spukspuginstruktion
flgezgdritterluv	subsidirereinsatz
flgezgdritterp7p9	takausinstruktion
flgezgdrltjetinklskaflieg	tellit
flgezglt	truppendiensttagestd
flugdiensteigenestraining	ttkwkflieger
flugdienstit	uegfdekf
flugunfall	uegflugdienstinstruktion
flugveranstaltungen	uplygone
jarinstruktionsflugd	v112
jartraining	v154
laaged	v211
ltjetvorbereitung	v46
lwosb	v91
meetingaarburg	verschiedenes
meetingaltenrhein03	wef2003
meetingaxalp03	wef2004
meetingbadragaz03	wef2005
meetingbarcelonae03	weitereauslandkampagnenrluv
meetingberlind04	worldeconomicforum
meetingbernbelp04	bmps
meetingbirrfeld04	pilausbberleben
meetingbratislava04	pilausbflugda3
meetingcognacf04	pilausbflugdpc7
meetingdaxf04	pilausbflugdtiger
meetingdecimomannui04	pilausbsat
meetingdittingen03	pilausbsras
meetingdiverserlw	pilausbtheorie
meetingemmen03	pilotenslw82
meetingevianf04	pilrs
meetingfairfordgb04	pilsi
meetinggenve04	pilsii
meetinggruyre03	piluosi
meetingkoksijdeb03	piluosii
meetinglangenthal03	plsaffacapilwxxijar
meetinglugano03	takausausbildung
meetingluganoagno03	ukfa18
meetingpayerne03	ukth98
meetingpayerne25403	umschulungen
meetingpayerne260903	wau
meetingrapperswil03	wsis03
meetingraron03	
meetingsalondeprovencef03	
meetingsamedan04	
meetingstmoritz	
meetingststephan04	
meetingtwenthenl03	
meetingverbier04	
meetingvolkeln104	
meetingvolketswil04	

Extraduties

Variablename	Variablename
ausbildunginglg	lobe
ausbildungmfs	luftfahrthindernisdienst
ausweiterbildung	luftraumbewirtschaftung
ausweiterbildungeigene	lufttransportdienst
flgllw	luftwaffexxi
flgtlgfdkfl81	lwseminar
teilnehmerinfgt	mitarbeitinfachteams
weiterbildungskurse;	nalouette3
air04pay	namraam
allgstabsarbkoor	ndauswertung
anlssebesuche	nderungsdienst
aocairoperationcenterjfacc	ndo27dornier
arbeitenfrdivag	neueirlwfll
arbeitenfrskyguide	nf5eftiger
assessment	nfa18cdhornet
ausbildungsgrundlagen	nflzsimtiger
ausbildungskonzepte	npc6
auskunftsert	npc7
banflugwinkelgert	npc9
bdatalinkfa18	npilotenausrstung
bearbeitungvonmilanfmph	nprojcbtth98
berichte	nsimhoupdate
bersetzungen	nska350c
bflintefislw	nsuperpumath89
bflorako	ntwinotter
bmehrzweckhelikopter	oeffentlichesamtstd
bmiltrspflz	osze
bmppersonalfhrungplanung	personaladministration
bneueirlwfll	personalbetreuung
bpilotenausrstung	personalfhrungberufsorg
bth98	personalplanungfhrung
dialoginstruktoren	personalplanungueg
dialoglobe	pfp
dienstleistungenfrdenstabkdt	pilotenausrstung
dienstleistungfrflbr31	posteinundausgang
dlfraal	projektelite
dlfrbaalw	projektom
dlfrbablw	projektsmds
dlfrbr31	protokolle
dlfrbr33	referatebeimilitrischenanlssen
dlfrdritteausserhalblw	referatebeizivilenanlssen
dlfrei	reglementrmfd95
dlfreistablw	remove
dlfrflbr31	sachbearbeitungflir
dlfrkdoueg85	sachbearbeitunglandestopograph
dlfrlvbfl31	saradministration
dlfrueg	schiessleiterluftboden
dlfrugop	schiessleiterluftluft
dlfrvorgesetzten	schulunglkf
doktrinforschungundentwicklung	sekretariatsarbeiten
doktrinverbreitung	sitzungenrapporte
eignungsabklrungfvs	softwareentwicklung
eignungsabklrungpil	softwareentwicklungfreigenenbed
einsatzbosnieneufor	stellungnahmen
einsatzkosovo	stellungnahmenberichte
entwveinsatz	stellvertretung
eufor	superuser
fachdlk	telefonausknfte
fachpublikationen	tglicheopeinsatzleitungeinsatzo
fas	tglicheoperationeneinsatzleit
feuerlscheinsatzportugal04	truppenantrge
fhrungorganisation	truppenversuche
fhrungsreglbsvorschr	untersttzungoperationenlw
fhrungssimulatoren	upgrade21u21
fislwfhrungssystem	usersupport
flowflplkdo	v143
flugdienstleitungcfo	v172
flugdienstleitunginklcfo	v24
flugsicherheit	v35
flugsicherung	v76
fotoundbildmaterial	v77

Source: own listing based on data provided by the Swiss Air Force.

Curriculum Vitae

Beat Hedinger

Address: Vogelsangstr. 9, 8006 Zurich, beathedinger@bluewin.ch

Date of Birth: 24.02.1976

Citizenship: Swiss

Civil status: Married

Education: Doctoral Studies at the University of Zurich,
Graduation: summa cum laude, December 2007

Studies of Economics at the University of Zurich,
Graduation: summa cum laude, November 2004

Swiss Air Force Pilot Diploma, May 2000

Airline Transport Pilot Licence, March 1999

Flight Instructor Licence, May 1998

Matura Type A (ancient languages), January 1996

Work: Swiss Air Force: Chief Flight Operations SPHAIR and
Executive Board Member SPHAIR, since July 2007
Swiss Air Force: Chief of academic training for pilots 2004-2007
Swiss Air Force: Helicopter-Pilot and Instructor since September 2000

Horizon, Swiss Flight Academy and Motorfluggruppe Zurich:
Civil Flight Instructor since 1998

Work abroad: Portugal and Greece: Swiss Air Force Fire Fighting 2004 and 2007
Balkans: Several Peace Support Operations 2003-2007

Philippines, Manila: SGV (Arthur Anderson & Co) and FE Zuellig
6 month Internship, spring 1996